LIKE TWO PIS IN A POD: AUTHOR SIMILARITY IN THE ANCIENT GREEK CORPUS

A Thesis

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Master of Science

by

Grant Justin Storey

May 2019

ABSTRACT

One commonly recognized feature of the Ancient Greek corpus is that some later texts imitate and allude to model texts from earlier time periods, but analysis of this phenomenon is mostly done for specific author pairs based on close reading and highly visible instances of imitation. In this work, we use computational techniques to examine the similarity of a wide range of Ancient Greek authors, with a particular focus on similarity between authors writing many centuries apart. We represent texts and authors based on their usage of high-frequency words to capture author signatures rather than document topics. We propose the Jensen-Shannon Similarity metric for measuring similarity between authors and show that it outperforms other common metrics for vector comparison. We then use this similarity metric to analyze author similarity across distances in time, finding high similarity between specific authors and across the corpus that is not common to all languages. We analyze these similar author pairs more closely and find the similarity is the result of similar usage of many different words rather than just a few.

BIOGRAPHICAL SKETCH

Grant Storey was born and raised in Berkeley, California. He was first introduced to computer science through a summer course on Java in 2008 and has been using programming to solve problems ever since. He received a Bachelor's Degree in Computer Science from Princeton University in 2017. While at Princeton, he worked with Professor Christiane Fellbaum on analysis of dialect and meter in Ancient Greek poetry.

After Princeton, Grant came to Cornell University, where he is graduating with an M.S. degree from the Department of Computer Science in May 2019. During his time at Cornell, Grant had the pleasure of continuing to apply computational techniques to the study of Ancient Greek texts while working with Professor David Mimno.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor David Mimno for all of his help at every step of this thesis. He pointed out the interesting phenomenon at the heart of this paper to get me started on this work in the first place, and whenever I was stuck on some seemingly impossible problem he helped come up with a workable solution. He was incredibly supportive and enthusiastic throughout the process, and I could not have completed this thesis without him.

I would also like to thank my minor advisor, Jeffrey Rusten, for his help with the classics side of this work. Our discussions were immensely useful both for figuring out what interesting questions to ask and for avoiding claims that do not make sense from the classical perspective.

Many thanks to Michael Weiss and Joshua Katz for additional help with sources on various features of Ancient Greek.

Lastly, a big thank you to Russell Peck for graciously letting me use texts from the TEAMS series for this research.

To my parents. Thank you for all your love and support.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Traditional analyses of the texts of Ancient Greek authors often discuss the earlier models imitated by an author. Arrian, writing in the second century CE, is said to have based his *Indica* on the model of Herodotus [20], and Apollonius Rhodius crafted his *Argonautica* in the style of Homer. As Antonios Rengakos puts it, the *Argonautica* is full of "imitations of Homeric phrases, verses, motifs or scenes and reproduces lexical, morphological, syntactical and metrical peculiarities of the old epic" [43]. Most prior analyses of this imitation examine highly visible, marked imitation, like the use of hapax legomena, words that appear once in the entirety of the Homeric corpus [32, 1], or the reuse of nearly identical phrase structures from earlier speeches [26].

Instead of focusing on a few specific instances, similarities between authors could also be analyzed using computational techniques. Recent work has begun to use computational methods to analyze classical texts, including authorship and allusion in Latin texts [30, 49] and the syntactic style of Attic prose [6, 7]. Our goal in this work is to analyze similarities in the writing style of Ancient Greek authors. The definition of "style" is a thorny problem that we cannot fully address within the context of this work, but at a high level we might expect the "style" of a section of text to be informed by some combination of its genre (e.g., fantasy, biography, epic poem, military history, philosophical treatise, etc), dialect (e.g., American vs British English), time period (English as a language changes from Shakespeare to Queen Victoria to J.K. Rowling), register (exquisitely highfalutin or simple), and other internal tendencies of the author. Parts of this "style" might vary across different works from the same author, or even possibly within a text in the case of a work with multiple styles like Faulkner's *As I Lay Dying*. We will

be focusing on the features of an author's writing that are more or less the same across all of their works and differ from author to author, which we will call an "author signature." While there are a variety of potential caveats and issues which we will discuss as they come up, the idea of an author signature provides us with a slightly more concrete starting point for comparing different authors.

We will attempt to capture author signatures using methods that have been successful in non-classics work on stylometry by focusing on authors' usage of very common words. The intuition behind this technique is that an author's usage of these words (say, the ratio between their usage of "but" and "and") is mostly unconscious and provides a signature (or fingerprint) for that author, while their usage of content words (say, "tropical" or "lodestar") reflects the topic of their work or conscious decisions about their style rather than a fundamental characteristic of the author.

While prior works focus on using author signatures of this sort to resolve authorship questions about a given text, our work is focused on comparing word usage of *known* authors writing in different styles and time periods. There is evidence from antiquity that authors writing in Ancient Greek frequently imitated the styles of earlier models and that individual authors had their own writing styles distinct from other authors. There has also been prior work on examining imitation with regards to content words, dialect forms, or salient phrases [45, 11] as well as analyses of allusion and intertextuality based on phrases or a small number words [12, 16]. However, to the best of our knowledge previous study has focused on comparing relatively short phrases between a few works, rather than considering "style" or author signature in entire texts across a larger corpus. Do Ancient Greek texts show signs of imitation when considering the most common words used for authorship analysis? How similar are Ancient Greek authors from the same time

period or the same genre? If we see later authors writing like earlier authors, *how* do they do it?

We make four major contributions in this work. In Chapter 2 we perform an initial exploration of the data, analyzing interesting features and preliminary discoveries to show that a feature set based on top words *does* capture author information that matches with prior scholarship. In Chapter 3, we explore various metrics for measuring similarity between authors and find that a similarity metric based on Jensen-Shannon Divergence detects similar author pairs better than other commonly used metrics while allowing analysis of the importance of individual word usage. Next, in Chapter 4, we analyze the relationship between author similarity and relative composition date, and show that Ancient Greek authors are often very similar to earlier models, even when writing centuries later. We find surprising similarity between authors both at the individual author level and across the corpus as a whole, yet this remarkable similarity is not a feature common to every language. Then, in Chapter 5, we take a closer look at word usage by authors who are similar despite writing at least four centuries apart, showing that the most similar author pairs achieve this similarity through their usage of most high-frequency words rather than imitating the usage of a few words exactly. Finally, we make concluding remarks in Chapter 6.

CHAPTER 2

DATA EXPLORATION

2.1 Data

2.1.1 Greek

Our analyses focus on texts available through the Perseus Digital Library's Greek Collection [48]. This corpus includes 466 works from 92 authors spanning from circa the 8th Century BCE to the 6th Century CE. Where possible, we divide each work into smaller segments: for example, we divide Herodotus' *Histories* (a single "work") into its nine books (each of which is a "segment"). When a work can not be broken up naturally, it is considered a single segment, so Euripides' *Medea* is both a work and one segment. This leads to 1,339 total segments. Of these segments, 1,139 (from 65 authors) are prose and 200 (from 27 authors) are poetry. The data is preprocessed to remove non-Greek characters and punctuation. Where possible, elided tokens are restored, so παρ' is restored to παρά. The full dataset has 9,709,776 tokens (total words) and 486,456 types (unique words) after preprocessing.

We examine the dataset in two forms. First, we group works by author, including all authors, no matter how little text there is. Each author has at least 2,000 tokens except for Bion of Phlossa (1,803 tokens) and the anonymous author of the fragmentary *Hymn to Dionysus* (just 144 tokens). The small sample size does not seem to adversely affect analysis of Bion. The short length of the *Hymn to Dionysus* does have an impact on the analysis, which we discuss below.

We also analyze the texts divided into individual segments, only considering segments with at least 1,000 tokens. This gives 1,204 segments to analyze (out of

the total of 1,339).

One of the major distinctions within this corpus is the difference between texts that are prose and poetry. Ancient Greek prose was subject to a set of constraints familiar to any author: it had to be grammatically correct, it had to convey ideas, and it was perhaps constrained by some cultural considerations about good writing practices. Ancient Greek poetry, in addition to those constraints, also had to conform to some poetic meter. There were a variety of meters used in different contexts (e.g., Dactylic Hexameter was the meter of epic poetry) but all of them mandated some pattern of long and short syllables in each line of poem. This means that, unlike prose, all Ancient Greek poems are naturally structured around lines and have extra constraints on how sentences can be constructed. Because of this, we expect that the language of Ancient Greek prose and poetry will show clear distinctions in our data.

There is one other potential issue for this data. Most work on stylometry is concerned with modern authors, so it is safe to assume that the published version accurately represents the author's original work and, by extension, their author signature. For this work, it is not so safe to assume that the modern print editions perfectly reflect the texts as first written by their Ancient Greek authors. Because the texts were written so long ago, the modern editions are actually an editor's reading of a set of medieval manuscripts, which may have long histories themselves. For example, the text of Euripides' *Medea* as Euripides wrote it was passed down by actors in Athens, then standardized by scholars in Alexandria, then transmitted in a variety of manuscripts through the medieval period to the modern day, where the manuscripts were combined by an editor into the single version of the text which we use. This has the potential to introduce differences between the modern text used in this work and the text as first written by Euripides. The modern

texts may have sections of the original text repeated by an overly dramatic actor or removed by an overly zealous editor, and they may have changes introduced by editors who thought a form was incorrect or scribes who made errors while copying the texts. Previous work on medieval Dutch texts has even shown that copying scribes can introduce their own signature to texts [29, 55]. We recognize that these are potential limitations of the current approach. However, accounting for them would require an extensive study of the interaction between the editorial and scribal artifacts of the manuscript tradition and the output of our method, which is beyond the scope of this work.

2.1.2 English and Icelandic

For comparison to Greek, we will also examine English and Icelandic texts. We expect English to show significant differentiation (from "Middle" to "Modern") over the past millennium, so it will provide a good benchmark for a language with major changes [39, 17]. English is also familiar to the English-speaking author. The English corpus has 206,546,514 tokens and 621,810 types. It is a combination of the following corpora, resulting in 166 authors with 2,759 unique works and 2,960 segments:

- Modern English texts from the Gutenberg Dataset [34], with a few duplicate texts and text with a mixture of prose and poetry removed.
- The plays of Shakespeare from the Shakespeare Corpus [46].
- Middle English texts from the TEAMS Middle English Text Series [50] supplemented by the Morte D'Arthur [37] and Canterbury Tales [10].

Many of the Greek texts have standardized spelling and morphology even across long distances in time, so we chose to examine Icelandic because it also has versions of older texts that have been updated to modern spelling and morphology. The modernized spelling in the Icelandic corpus is slightly different from the more conservative spelling of the Ancient Greek corpus, but the important feature is that in both cases the spelling is standard across all the texts. We hope that Icelandic will therefore provide a good comparison of a language with standardized orthography and morphology across a long period. In addition, beyond the standardization of spelling in our corpus, Icelandic is considered to be a relatively conservative language in terms of change over time [39, 17, 2]. The Icelandic corpus has 7,591,751 tokens and 290,926 types. It is a combination of the following corpora, resulting in 196 authors with 213 unique works, each of which is one segment:

- Icelandic Sagas from the Saga Corpus, with duplicate manuscripts removed [44].
- The Icelandic Parsed Historical Corpus (IcePaHC), a collection of texts from 1150-2008 CE [25]. Duplicate manuscripts and translations were removed.
- 21st Century Icelandic texts from the Tagged Icelandic Corpus (MÍM) [21]. This corpus contains many kinds of texts, but we only used the subset of the texts labeled books, with articles by Baldur Jónsson and recipe books removed.

2.2 Document Representations through High-Frequency Words

When considering authors and segments, we need a way to represent them so that they are easy to compare while still maintaining information about their author signature. We do this by representing each author as a set of individual **features**, each of which is some piece of information about the author or segment. All of the features together is called the **feature set**. Our choice of feature set is very important, as maintaining too much information (say, the full list of words in the text, in order) makes it hard to compare texts easily, while using too little information, like only the name of the author or the first word in a segment, would mean that comparisons are not particularly useful.

In this work, our feature set for measuring author signatures is the frequency of the most frequently used tokens across the corpus. Usage of the most frequent words has shown promising results for identifying author signatures in the past, in particular because content words are more dependent on genre and topic matter [8]. Some work has found that frequent words alone can do better than part of speech information or a combination of the two [53]. Ancient Greek is highly inflected, so a common first step when working with these texts is to replace surface forms with lemmata.[1] However, in this work we consider only surface forms, not lemmata, for two reasons. The first reason is that, when considering the most frequent tokens, there is valuable information in the inflection of words: for example, an author's usage of τῆς as compared to τόν is an interesting and potentially relevant distinction. While lemmatization is useful for grouping together different forms of content words, in this work we are not considering content words and thus

[1] In English, this process would involve replacing forms like "trees" with "tree" and "held" with "hold."

lemmatization is less essential. The second reason for not using lemmatization is that the lemma for many tokens is ambiguous and it would require concerted manual work to resolve these ambiguities across 9.7 million tokens for a reliable analysis.

Where possible we use a list of the top 145 words across all texts combined with the list of the top 100 words in only the poetic texts. Due to overlap between these two sets, this yields a list of 172 total words. The full list of tokens used can be found in Appendix C. These word cutoffs were chosen because lists with larger cutoffs included words that were very frequent in a few texts but not generally applicable. For example, the 146th most common word is the name Σωκράτης (Socrates), which does not occur at all in over half the texts. Common words from poetry were included because the word usage in Ancient Greek poetry and prose has key distinctions, as we will see below in Section 2.3. Since the corpus is dominated by prose, including more words that are specifically relevant to poetry (including more poetry-specific words like κεν and words that appear more often in poetry than prose, like Ζεύς) helps better capture the signatures of poetic texts and ensure differences in their word usage are detected. In some places below we also use only the top 145 words; for example, when making comparisons to the Icelandic corpus, which has no poetry.

In the following, our feature set is the frequency of the top words within each author or segment. We therefore represent each author and segment with a vector P consisting of 145 or 172 features, where P_i corresponds to the frequency of word i within the given author or segment.

$$P_i = \frac{(\#\text{ of occurrences of word } i \text{ in text } P)}{(\text{total } \#\text{ of words in text } P)}$$

Note that the total number of words includes *all* words, not just the top 145/172. When we consider two texts at once, we label the first P and the second Q.

2.3 Detecting Author Characteristics

Before using the feature set of most frequent words to analyze author similarity, it is important to check whether this feature set actually captures information about the texts. It is certainly plausible that reducing each text to only its usage of the 145 or 172 most common words would lead to something useless.[2] If the top words do not seem to carry any meaningful information about the texts, we should not try to measure similarity with them. To get a sense for what kind of information an author's usage of the most frequent can tell us, we analyze the Ancient Greek data as a whole to see if there are any interesting properties that can be detected just from the word representation vectors. Does the data allow us to clearly differentiate authors based on characteristics like genre, dialect, or time period?

2.3.1 Visualizing the Data

First, in Figure 2.1, we create a two-dimensional tSNE projection [36] of the texts based on the *top 145* words (without any words from poetry specifically). The tSNE visualization takes the points from a very high-dimensional space and maps them in two dimensions, preserving relative distances as much as possible, so points that have similar word usage appear closer together here. The most salient feature of Figure 2.1 is the large gap between authors writing prose (red) and poetry (blue), which shows that there is a clear difference between poetry and prose authors based on how they use the most frequent words. It is important to note that

[2] e.g., that sentence would be "It is that to only its of the or most would to." Claiming this is all the information about the sentence we need to determine an author's signature without any evidence would be unwise.

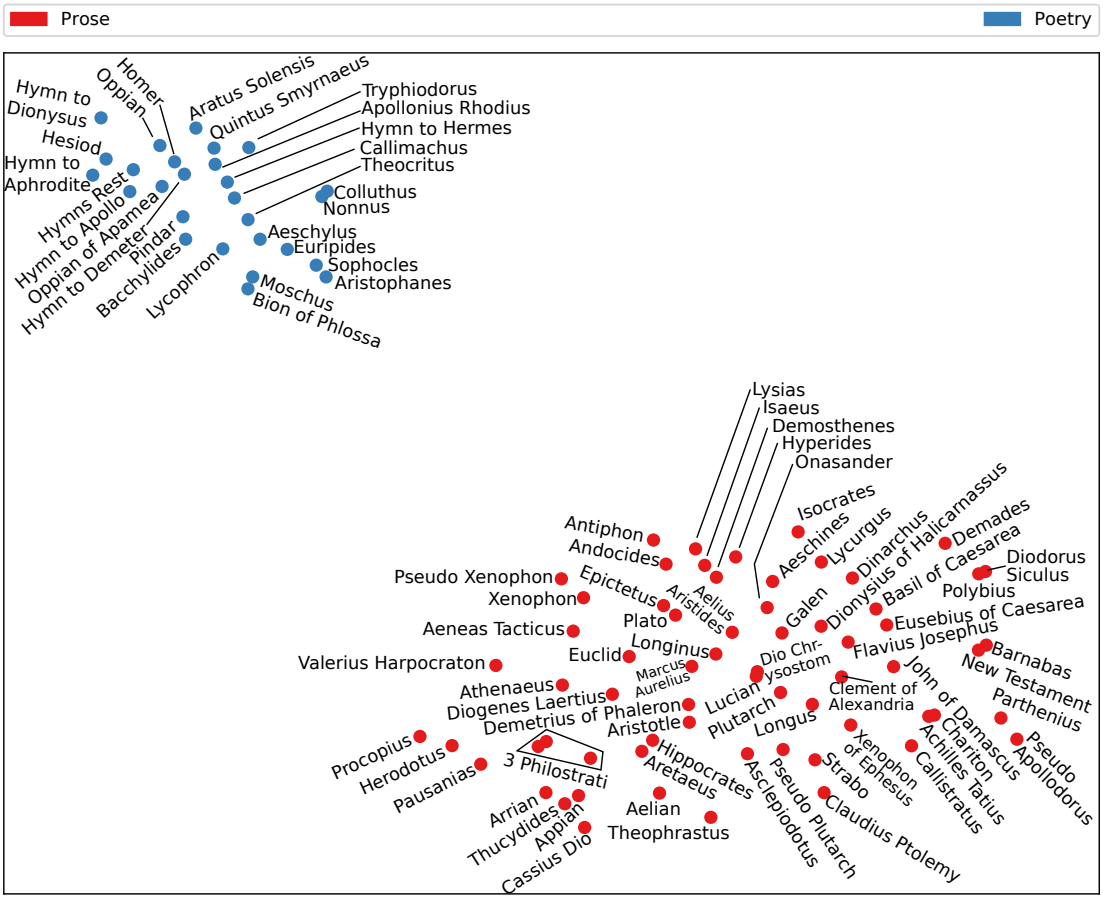

Figure 2.1: A two-dimensional tSNE Projection of the authors based on their usage of the top 145 words, with similar authors grouped together. Points are clustered without knowledge of their genre, so the separation between poetry and prose is based entirely on different word usage in the two genres.

the algorithm clusters points based *solely* on their usage of top words, so it found this difference between prose and poetry without knowing which authors wrote in which genre. This chart shows that there is a clear difference between prose and poetry word usage — in fact, the difference is so clear that the *Hymn to Dionysus* is clearly recognized as poetry based on a tiny text sample (144 total tokens). Because the corpus is dominated by prose, in order to help capture different word usage within poetry texts it therefore make sense to augment the list of top words with words that are common in poetry specifically.

The clear distinction between prose and poetry texts gives some hope that other characteristics of these authors might be distinguishable as well. More specifically, for any given characteristic we consider three broad possibilities:

A. The different categories for this characteristic (e.g., poetry and prose) are so distinct that they form entirely separate clusters in the high dimensional space which are reflected as separate clusters in the tSNE visualization.

B. The different categories can be distinguished based on the given features (e.g., by using simple machine learning) but are not so distinct that they form independent clusters.

C. The different categories are not easy to separate based on the given features and distinguishing between them is either impossible or would require advanced machine learning (within the scope of this section we do not differentiate between these two possibilities).

If every category of interest matches hypothesis C, our set of features will probably not be useful for capturing information about author signatures.

In Figure 2.2, we show clustering based on the top 172 words (including both top 145 words and additional poetry words), with four different colorings, based on poetry vs prose (upper left), a more narrow genre distinction[3] (upper right), time period (bottom left) and dialect (bottom right). While poetry and prose form independent clusters, none of these other categories shows a clear distinction. This means that the poetry/prose distinction matches hypothesis A above, but we can rule out hypothesis A for a more narrow account of genre, time period, and dialect.

[3]between Romance, Speeches, Military/Historical Prose, Judeo-Christian Prose, Philosophy, Other Prose, Comedy/Tragedy, Epic Poetry, Didactic Poetry, and Other Poetry.

Figure 2.2: tSNE Projection of authors based on their usage of the top 145 words and top 100 poetry words, with similar authors grouped together. The four charts show the same authors colored on different criteria: prose and poetry (top left), more narrow genres like military prose and epic poetry (top right), time period (bottom left), and dialect (bottom right).

We also briefly consider the possibility that the tSNE projection is dominated by the difference between poetry and prose and therefore fails to pick up on other distinctions. To try to adjust for this, we normalize the data to remove the distinction between poetry and prose. First, we find the "average" poet and "average" prose author by taking the mean of the frequencies across all authors in the two categories. Then, we calculate a normalized feature vector for each author by subtracting the feature vector of the "average" author of their genre from their feature vector. This should hopefully remove the clear distinction between prose and poetry authors as a whole while still preserving other differences. This would roughly correspond to modifying the original texts by removing some of the more common prose words from the prose texts and more common poetry words from the poetry texts, but since we are working with feature vectors instead of actual words at this point, we sometimes end up with artifacts like negative frequencies that could not occur in real texts.

The projections under this scenario are visible in Figure 2.3. It is clear from the upper left that this analysis removes the distinction between prose and poetry as groups, but we still do not find a clear distinction between more narrow genres, time period, or dialect. There is some meaningful grouping at a local scale: many similar groups of authors appear together, including the playwrights of Athens and the 3 Philostrati, and we see some clustering of similar author pairs using the same dialect in the bottom right, but overall there is no clear clustering of different groups, and the local groups were present in the non-normalized data as well. So at a first glance the top words do not appear to show clear distinction between time period or dialect.

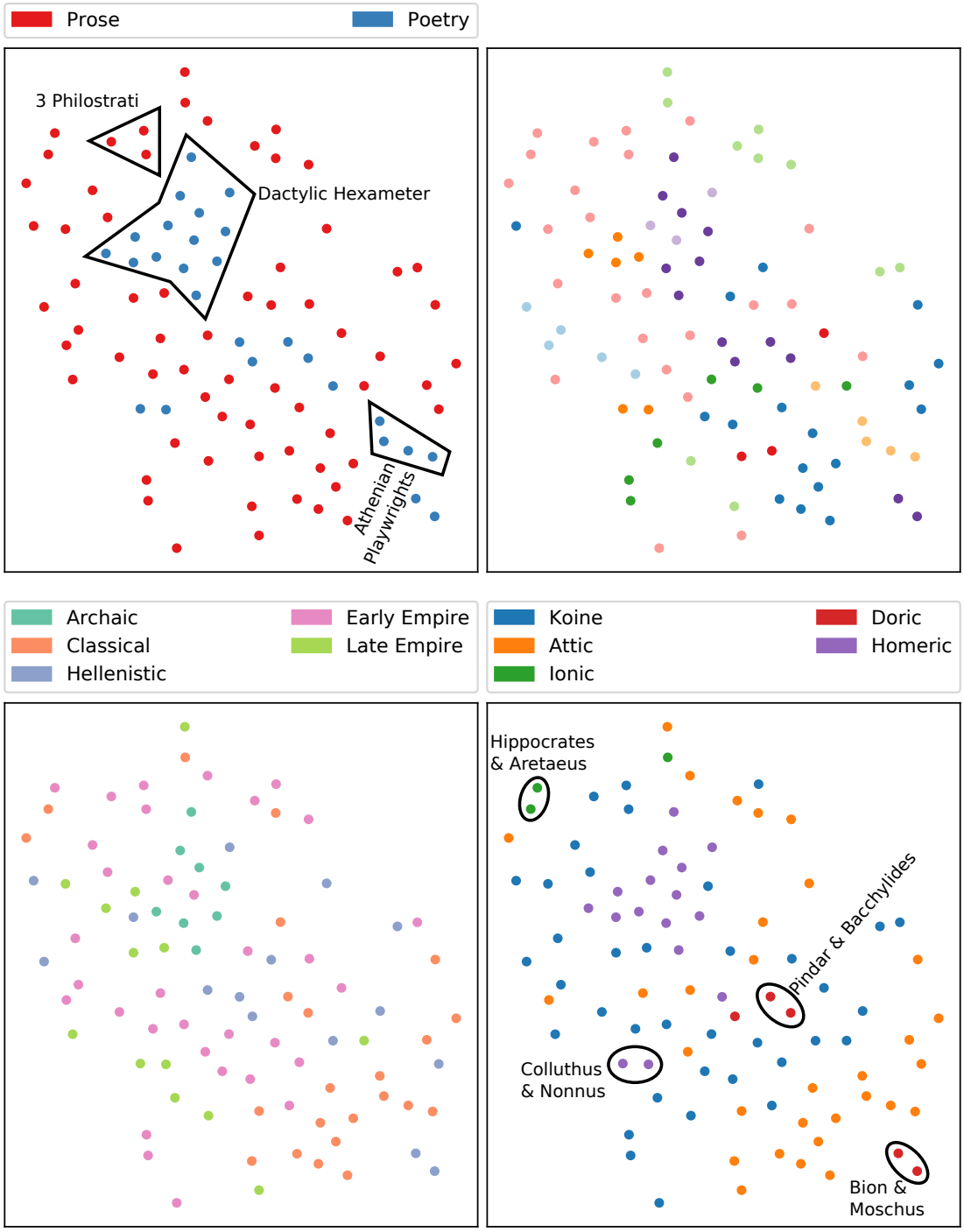

Figure 2.3: tSNE Projection of authors based on their usage of the top 145 words and top 100 poetry words, with features normalized by poetry and prose. The four charts show the same authors colored on different criteria: prose and poetry (top left), more narrow genres like military prose and epic poetry (top right), time period (bottom left), and dialect (bottom right).

Prediction Task	Majority Class	KNN	Naive Bayes
Genre of Authors	0.706061	**1.000000**	**1.000000**
Dialect of Authors	0.401010	**0.739394**	0.725253
Time Period of Authors	0.357576	**0.486869**	0.435354
Genre of Segments	0.841326	0.998335	**1.000000**
Dialect of Segments	0.451003	**0.950997**	0.766606
Time Period of Segments	0.485810	**0.945174**	0.644503
Author of Segments	0.104646	0.847199	**0.872922**

Table 2.1: Results of running simple machine learning on the frequency data based on the top 145 words plus top 100 poetry words.

2.3.2 Author Characteristic Prediction

The unsupervised tSNE Projections do not show a clear distinction between categories beyond poetry/prose, so we can rule out hypothesis A for these other categories. In order to determine whether hypothesis B or C is correct, we run three simple classifiers on the data: Majority Class, K Nearest Neighbors (with K=2 chosen from the set {1, 2, 3, 5, 10, 20} based on best performance on the training folds.) and a Multinomial Naive Bayes.[4] For each classifier, we divide the data into nine folds.[5] For each fold we evaluate a model trained on the other eight folds. The accuracy for each method reflects the average performance across all nine folds. Results are visible in Table 2.1.

Both the KNN and Naive Bayes classifiers do extremely well at predicting genre (poetry vs prose) of authors and segments, achieving >99% accuracy in all cases. Dialect and time period prediction for authors are slightly worse, but still far better than the majority class baseline. When considering individual segments, K Nearest Neighbors is actually very predictive of dialect and time period, achieving around 95% accuracy. Based on the performance of KNN on guessing authorship

[4]See Appendix A for brief explanations of how these work.
[5]Since there are 92 authors, this provides around 10 authors to test on for each fold.

of segments, it appears that in the majority of cases the nearest neighbor of a segment is another segment by the same author, but this technique shows additional improvement beyond just guessing the author.

The final row of Table 2.1 also shows the accuracy of these classifiers at predicting the author of given segments. Naive Bayes accurately classifies 87% of segments, with KNN performing almost as well. This performance is not perfect and could certainly be improved, but it does show that the features are at least partially predictive of authorship, even with relatively simple techniques.

When considering the authors as a whole, the data can be used to predict dialect and time period with some accuracy, but not particularly well. With the larger amount of data present in the segment-by-segment analysis, dialect, time period, and to a lesser extent author are all able to be determined with high accuracy. So at the author level, dialect and time period fall under hypothesis C above, while at the segment level they are reasonably predictive, even if they do not show clearly distinct clusters, which corresponds with hypothesis B.

All of these results could perhaps be improved by using more complex classifiers with greater hyperparameter tuning, but that task is beyond the scope of this paper. Even these results show that the feature set we have chosen — examining the frequency of the top words within an individual segment or an author's work as a whole — captures information about these texts including genre (poetry vs prose), and, when there are a large number of samples in the segment case, dialect, time period, and authorship.

2.4 Analyzing Author Segments

In addition to analyzing how these features interact with characteristics of authors like their genre or dialect, there are also many ways to examine individual segments of authors. As with the authors as a whole, we can create a 2-dimensional tSNE projection of the many different segments. For the most part, segments by similar authors cluster together nicely, but there are some exceptions. Four of these exceptions are highlighted in Figure 2.4. In each of the four charts, a single author's segments are colored in blue, with one clear outlier.

In the upper left, we see speeches attributed to Demosthenes, with speech 59 (*Against Neaera*) distinct from the rest. Critics since Dionysius of Halicarnassus have considered this speech to be by an author other than Demosthenes [13], with some modern scholars attributing the text to Apollodorus [28]. On the other hand, the upper right shows that Isocrates' Speech 21, *Against Euthynus*, is also a clear outlier, but the attribution of this speech to Isocrates is quite secure [24]. While the text is almost certainly the work of Isocrates, it *is* recognized as being in a markedly different style.

The other two cases are slightly more complex. Xenophon's authorship of the *Cynegeticus* (bottom left) has long been both challenged and defended, but it is unquestionably written in a very different style from the rest of Xenophon's work [19]. Our analysis (bottom right) supports a recent study by Thomas Koentges that found Plato's *Menexenus* to be the most unusual of his works based on a few computational analyses [31]. This text has been suspected of being non-Platonic in the past, but not recently [52]. The causes of this difference deserve more exploration; without further analysis, the data we present cannot clearly distinguish between the "different author" and "same author, unusual style" possibilities.

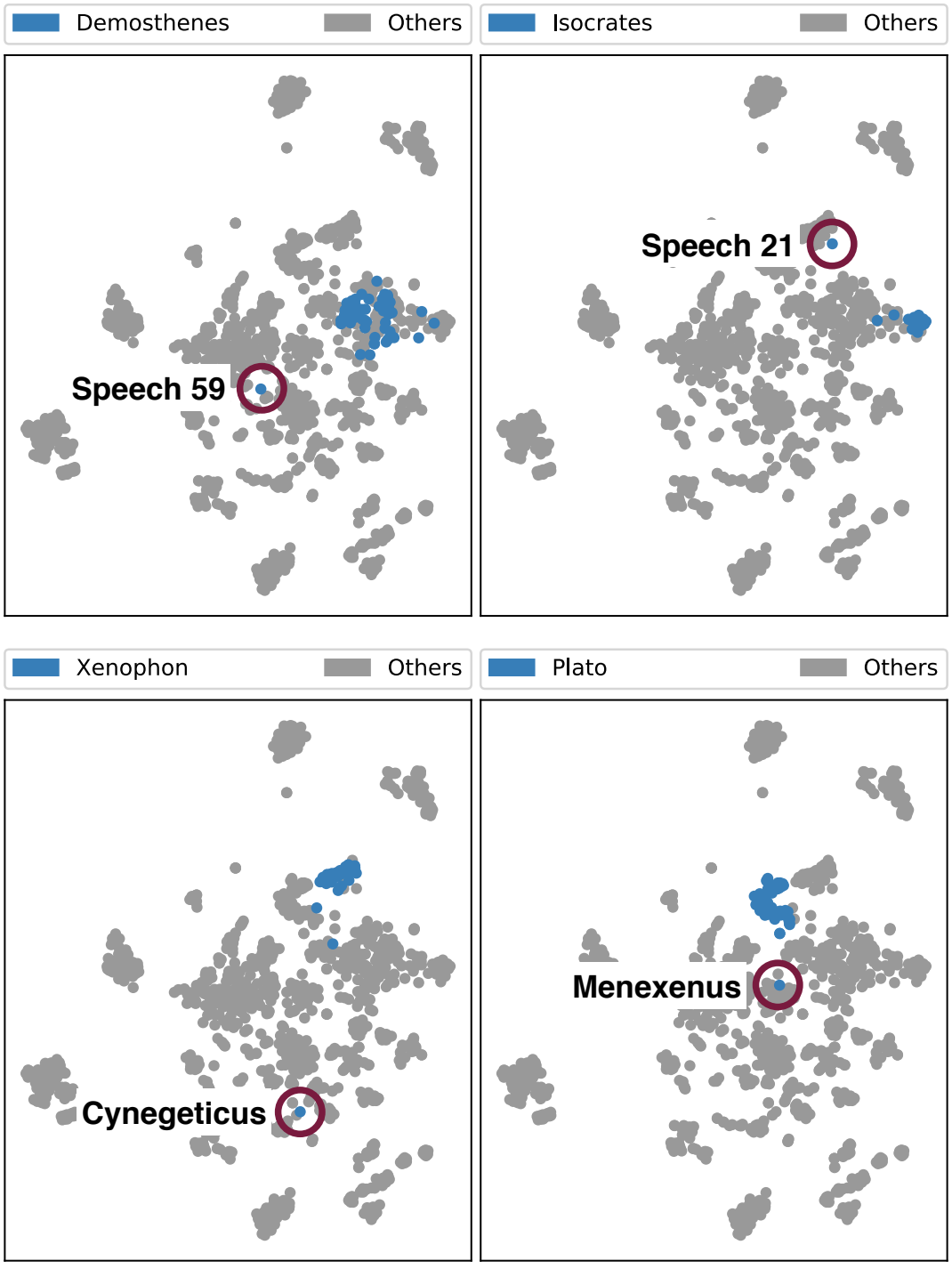

Figure 2.4: Four charts showing segments that are outliers for different authors. Segments written by the author highlighted in each subplot are blue, while all remaining segments from the Ancient Greek corpus are gray. Segments are grouped using a two-dimensional tSNE Projection.

Taking these four outliers together, we see that our feature set is detecting texts with incorrect attribution, texts of markedly different style, and texts flagged as unusual by other scholarship using slightly different methods. We do not claim to be able to solve any problems surrounding the authorship of these outlier segments, but using this set of top words to analyze text segments *is* capturing some aspect of an author's signature that tends to be different for different authors and works of dramatically different style. This feature set also produces results that agree with prior scholarship.

Across a few different data analyses, we have shown that our feature set does seem to be capturing valuable information about authors and texts, including to some extent information about genre, time period, dialect, authorship, and style. All of these analyses could be extended and refined to try to provide more definitive answers to the questions touched on or craft better classifiers, but the focus of this chapter is on validating our choice of features. Based on the performance of simple classifier with respect to various author characteristics and the agreement with prior work concerning unusual segments by certain authors, we can be confident that these features are capturing characteristics of the authors and will be a good basis for comparing varied writing signatures.

CHAPTER 3

MEASURING SIMILARITY BETWEEN AUTHORS AND SEGMENTS

In order to calculate the similarity between two authors or two segments, we would like to use a consistent metric. A variety of methods have been used in authorship attribution, including comparing raw frequencies [5], using Euclidean distance [33], and bootstrap consensus trees [54], among others. When choosing a metric for *this* work, the most important quality is that it captures the similarity between authors and segments as accurately as possible. In addition, there are a few other desirable traits for this metric:

- The similarity metric should be *symmetric*:

$$\text{Similarity}(P, Q) = \text{Similarity}(Q, P)$$

  So the similarity between Homer and Apollonius Rhodius should be the same as the similarity between Apollonius Rhodius and Homer.

- The similarity metric should decompose over individual words, that is, we should be able to tell how much each individual word (καί, δέ, etc) contributed to the similarity between two authors.

- The similarity metric should balance absolute and relative differences in the usage of words. This means the difference between 0 and 30 (absolute 30, relative 100%) occurrences should be more heavily weighted than the difference between 5,000 and 5,040 occurrences (absolute 40, relative 0.39%), but the difference between 0 and 2 occurrences (absolute 2, relative 100%) is not weighted more heavily than the difference between 1,000 and 8,000 (absolute 7,000, relative 77%).

Before examining each similarity metric, we would also like some way to get a sense for how a metric performs. To help get this intuitive sense for how each metric is comparing two texts, we set up a simple experiment. We start with book 1 of Thucydides' *Peloponnesian War* and books 1-6 of Nonnus' *Dionysiaca*,[1] which we expect to be quite different from Thucydides, as it is written in a different genre, dialect, and time period. We then replace tokens in the *Dionysiaca* sample one at a time with the token in the same position in *Peloponnesian War* Book 1 (e.g., we replace "κειμήλιον", the 2,000th word in *Dionysiaca* 1, with "ἐξηρτύετο", the 2,000th word in *Peloponnesian War* 1), creating a series of texts that are each one word closer than the previous text to the *Peloponnesian War* book 1. We use the given metric to compare *Peloponnesian War* book 1 to each of these individual steps, creating a graph that shows the performance of the metric as it compares texts that range from completely different to identical. See Figure 3.1 for an example.

3.1 Metrics Examined

We now examine five potential metrics for determining the similarity between two texts based on high-frequency words.

3.1.1 Cosine Similarity

The first metric we consider is cosine similarity, a common metric for comparing two vectors. Cosine similarity measures the angle between these two vectors, so identical vectors have a cosine similarity of 1 and opposite vectors have a cosine

[1] That is, a section of the *Dionysiaca* of the same length as *Peloponnesian War* book 1.

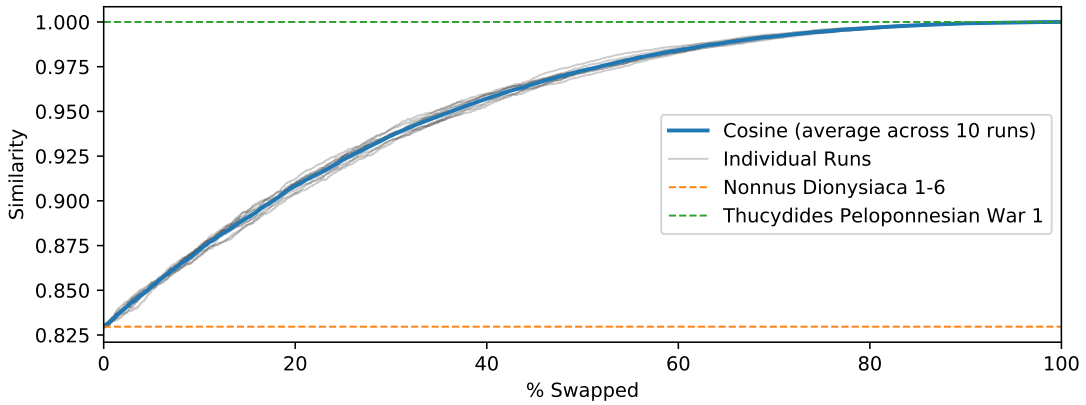

Figure 3.1: Cosine Similarity between Thucydides *Peloponnesian War* Book 1 and the first 6 books Nonnus' *Dionysiaca* with x percent of the words replaced by *Peloponnesian War* book 1 (averaged across 10 runs).

similarity of -1. However, since all of our vectors in this work have positive values, cosine similarities range between 0 and 1. Cosine similarity is calculated as follows:

$$\text{Cosine Similarity}(P, Q) = \frac{\sum_i P_i * Q_i}{\|P\|\|Q\|}$$

Where

$$\|P\| = \sqrt{\sum_i P_i^2}$$

To get a sense for how well this metric captures similarity, we show the results of the experiment described above for Cosine Similarity in Figure 3.1. As expected, the similarity between Thucydides' *Peloponnesian War* book 1 and the text increases as the text slowly changes from Nonnus' *Dionysiaca* to *Peloponnesian War* book 1, with some variance across different runs. The metric recognizes the reference text as basically the same as *Peloponnesian War* book 1 once the two texts are about 85% the same. This means that we can replace about 15% of the tokens in a work with tokens from a totally different text and cosine similarity will still be very close to one, so smaller cosine similarity values reflect a serious difference in texts.

While cosine similarity is a common technique for comparing vectors, it does

not allow easy comparison of the contribution of individual words to the similarity. Instead, one would have to rerun the similarity with that single word left out, then compare that change to the change when taking out each other word individually. This is rather time-consuming and not very intuitive, so we consider other metrics as well.

3.1.2 Manhattan Similarity

Manhattan distance (or City-Block distance) compares the absolute differences between the components of two vectors, the name coming from the fact it is calculating the city blocks between two points rather than the straight-line distance. It is a rather common distance metric for vectors, though less commonly used for comparing text frequencies. Manhattan distance is calculated as follows:

$$\text{Manhattan Distance}(P, Q) = \sum_i |P_i - Q_i|$$

And to convert this to a similarity metric, we say

$$\text{Manhattan Similarity}(P, Q) = 1 - 0.5 * \text{Manhattan Distance}(P, Q)$$

Since the values in the vector are frequencies, they sum to at most one, so the maximum distance is 2 and we must therefore normalize by 0.5 to get a similarity in the 0 to 1 range.

In Figure 3.2, we once again see a reasonably intuitive increase in similarity as the text transitions from similar to different, although the range in similarity is much larger than the range for Cosine Similarity (0.5 to 1.0 vs 0.825 to 1.0).

Manhattan Similarity is symmetric and decomposes over individual words, but it focuses on absolute differences to the exclusion of relative distances. That is, the difference between words with frequencies 1,000 and 1,030 (30) is considered

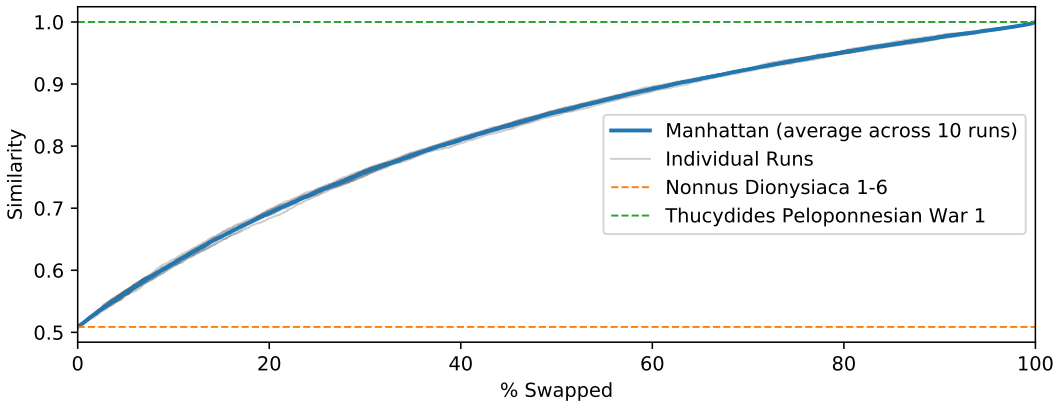

Figure 3.2: Manhattan Similarity between Thucydides *Peloponnesian War* Book 1 and the first 6 books Nonnus' *Dionysiaca* with x percent of the words replaced by *Peloponnesian War* book 1 (averaged across 10 runs).

more significant than the difference between words used with frequencies 0 and 25 (25), even though the second case likely shows a much more important difference in word usage.

3.1.3 Canberra Similarity

Canberra distance is similar to Manhattan distance but compares the relative differences instead of absolute differences. Canberra distance is calculated as follows:

$$\text{Canberra Distance}(P, Q) = \sum_i \frac{|P_i - Q_i|}{|P_i| + |Q_i|}$$

To convert this to a similarity metric, we use

$$\text{Canberra Similarity}(P, Q) = 1 - \frac{1}{\#\text{ of features}} * \text{Canberra Distance}(P, Q)$$

Since the values for each feature range between one and zero, we normalize by the number of features.

Figure 3.3 shows that this metric is a bit more volatile than the previous two. This is likely because a single occurrence of a word can have a major impact

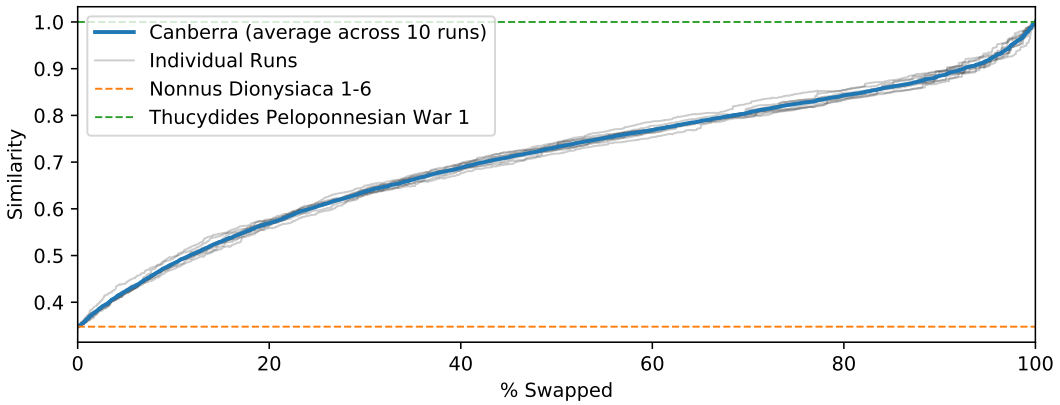

Figure 3.3: Canberra Similarity between Thucydides *Peloponnesian War* Book 1 and the first 6 books Nonnus' *Dionysiaca* with x percent of the words replaced by *Peloponnesian War* book 1 (averaged across 10 runs).

on the less frequent words due to the relative nature of the difference. When compared to the earlier metrics, this metric views the reference text as less similar to *Peloponnesian War* book 1 until it is almost 100% the same, so it makes more distinction between texts that are quite similar but not identical.

In contrast to Manhattan Similarity above, Canberra Similarity focuses on relative differences at the expense of absolute differences, but this means it has a different problem: when one author uses a word once or twice and the other author does not use it at all, it weights this difference very heavily. So instead of considering Manhattan or Canberra Similarity, we would like to examine a metric which balances absolute and relative differences.

3.1.4 Burrows' Delta

One of the most commonly used metric for authorship analysis is Burrows' Delta, which compares normalized relative usage of top words [8, 40, 3, 4, 14, 55]. One begins with a series of texts and calculates the mean μ_i and standard deviation σ_i for each feature i, then normalizes all of the data $P_i' = \frac{P_i - \mu_i}{\sigma_i}$. The similarity

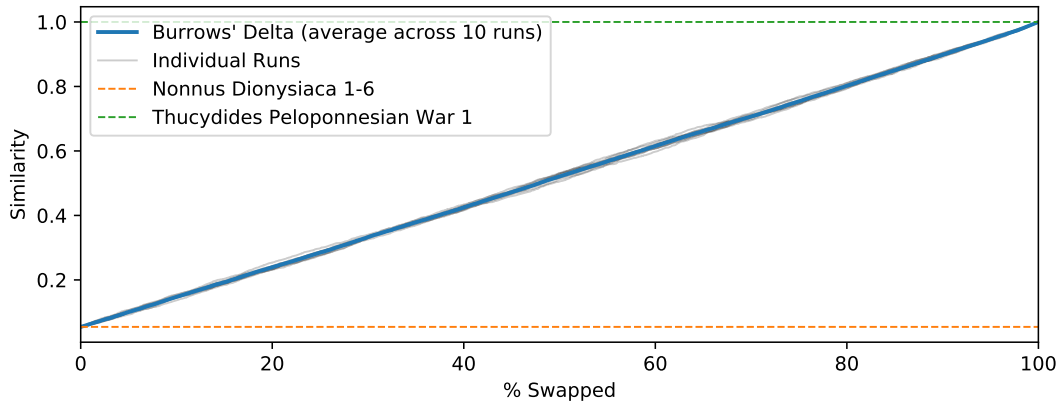

Figure 3.4: Burrows' Delta between Thucydides *Peloponnesian War* Book 1 and the first 6 books Nonnus' *Dionysiaca* with x percent of the words replaced by *Peloponnesian War* book 1 (averaged across 10 runs). Values have been normalized based on **all** segments.

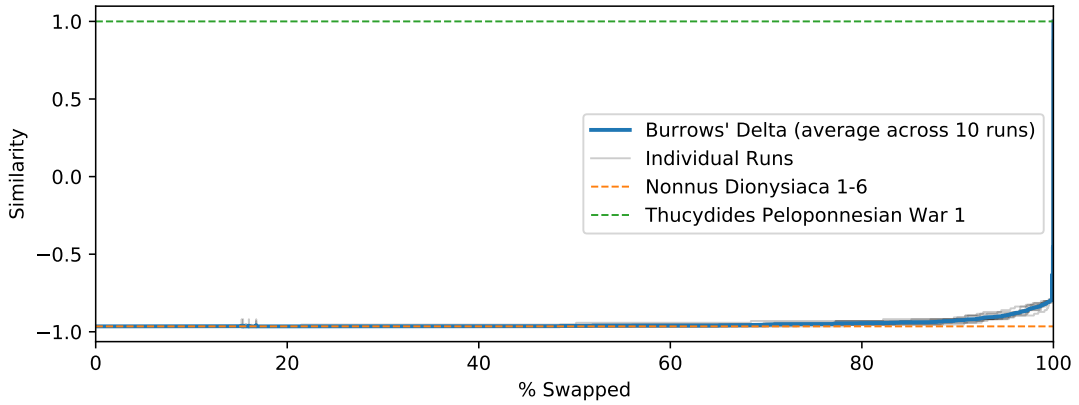

Figure 3.5: Burrows' Delta between Thucydides *Peloponnesian War* Book 1 and the first 6 books Nonnus' *Dionysiaca* with x percent of the words replaced by *Peloponnesian War* book 1 (averaged across 10 runs). Values have been normalized based on the two segments only.

between texts is then 1 minus the Manhattan Distance between the normalized features of the two texts. This does not necessarily range between 0 and 1, but more similar texts have higher numbers so it is still a valuable similarity metric.

Figure 3.4 shows that the metric captures a remarkably linear climb in similarity from about 0.05 to 1.0. This shows one nice feature of the metric, which is that with the appropriate amount of texts the most different texts have a score around zero and the most similar texts have a score closer to one. However, be-

cause Burrows' Delta is a metric designed for detecting the most likely candidate for authorship of a text [8], it normalizes the data using the corpus as a whole. This means that it is dependent not only on the two texts being compared, but also on the rest of the texts considered.

In Figure 3.4 we show the metric's performance when it compares these texts within the context of **all** segments, but if we only compare *Peloponnesian War* book 1 and the hybrid without using the rest of the corpus to normalize, we get a different picture, visible in Figure 3.5. This shows a rather volatile series of step-wise functions and does not show much improvement until the hybrid text is 100% the same as *Peloponnesian War* book 1. Because the normalization happens over just two texts, there are only two possibilities for each word: completely the same or completely different. As the frequency of a word switches from completely the different to completely the same, the similarity jumps up a single step. However, most words do not have identical frequencies until the texts are essentially identical.

Burrows' Delta does decompose over individual words and balances absolute and relative differences thanks to the normalization, so it meets our desired criteria above. However, since it was designed to choose the best candidate authors from a large set of potential candidates, it does not work well for comparing just two texts and is very dependent on the overall corpus used as the context for the comparison.

3.1.5 Jensen-Shannon Similarity

Zhao et al. found Kullback-Leibler Divergence (KL-Divergence) to be a useful metric for determining authorship of a text [57], but KL-Divergence has one serious

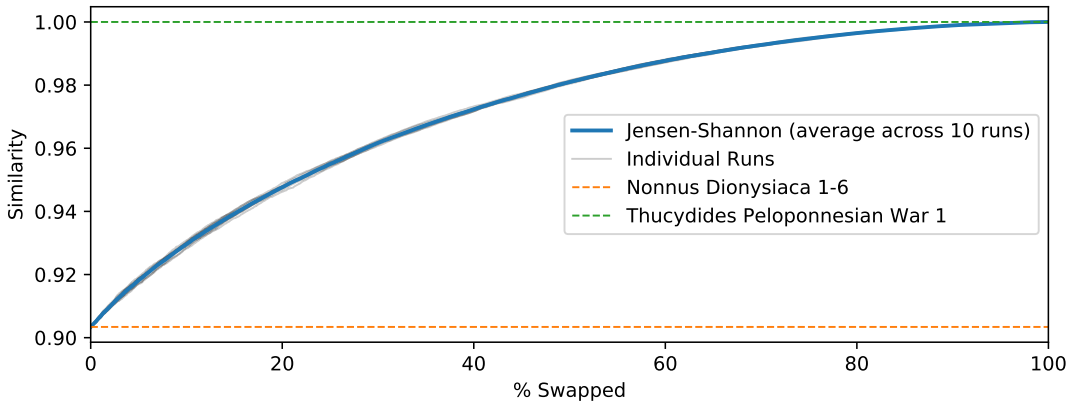

Figure 3.6: Jensen-Shannon Similarity between Thucydides *Peloponnesian War* Book 1 and the first 6 books Nonnus' *Dionysiaca* with x percent of the words replaced by *Peloponnesian War* book 1 (averaged across 10 runs).

flaw for this application: it is not symmetric. To address this issue, in this work we consider a metric based on Jensen-Shannon Divergence, a symmetric version of KL-Divergence. Jensen-Shannon Divergence is defined as follows:

$$\text{Jensen-Shannon Divergence}(P, Q) = \frac{1}{2}\left(\text{KL}\left(P, \frac{P+Q}{2}\right) + \text{KL}\left(Q, \frac{P+Q}{2}\right)\right)$$

Where KL is Kullback-Leibler Divergence,

$$\text{KL}(P, Q) = \sum_i P_i \ln \frac{P_i}{Q_i}$$

To get a similarity metric rather than a divergence, we calculate

$$\text{Jensen-Shannon Similarity}(P, Q) = 1 - \text{Jensen-Shannon Divergence}(P, Q)$$

Also, since this metric compares two probability distributions, the metric compares the counts of each of the top words plus the total number of non-top words. Plus-one smoothing is also applied to frequencies to prevent probabilities of 0.

In Figure 3.6 we see the expected smooth transition as the comparison text changes from similar to different. In addition, the spread of the ten individual runs is closer to the average than for most of the other metrics examined, showing that Jensen-Shannon has a bit more stability regardless of different sets of changing

words. So Jensen-Shannon Similarity appears at first glance to do an appropriate job of capturing the difference between two texts. Jensen-Shannon is symmetric, decomposes over individual words, balances absolute and relative distances, and also is not dependent on the corpus used, so it meets all of our desired criteria. But to determine which metric is most appropriate, we turn to a more concrete evaluation.

3.2 Metric Evaluation

In order to evaluate whether these metrics appropriately capture similarity between texts, we must compare the performance of our similarity metrics. First, we examine all segments that have another segment from the same work (e.g., *Iliad* book 1 has all of the other books of the *Iliad* in the same work) and determine whether the most similar segment is from the same work. Second, we compare all segments that have another segment by the same author, and determine whether the most similar segment is by the same author. The intuition here is that in general the segment most similar to a given segment should be from the same work and by the same author.

For each metric, we also examine its performance when using the top 145 overall words, top 145 overall words + top 100 poetry words (for a total of 172 words), and the top 172 overall words. Examining the top 172 words allows us to differentiate improvement due to using top poetry words from improvement due only to using more words. As a comparison for the significance of our metrics, we consider two baselines: Cosine Similarity, which is commonly used for comparing distributions, and Burrows' Delta, which is commonly used for analyzing authorship based on top words. For both of these metrics we choose the top 172 overall words as

Metric	Top 145 + Top 100 in Poetry	Top 172	Top 145
Jensen-Shannon	**91.33%†‡**	90.86%†‡	90.26%†‡
Burrows' Delta	88.84%†	87.89%	88.00%
Manhattan	87.89%	87.53%	87.41%
Canberra	88.48%	87.17%	86.70%
Cosine	86.22%	86.22%	86.94%

Table 3.1: How well similarity metrics based on a given set of words identify whether two segments come from the same work.
†: Results very significant ($p < 0.01$) when compared to Cosine (172).
‡: Results very significant ($p < 0.01$) when compared to Burrows' Delta (172).

our baseline to determine whether using extra poetry words shows improvement. These results are visible in Tables 3.1 and 3.2.

When comparing segments from the same work (Table 3.1), Jensen-Shannon Similarity with extra poetry words does the best job of identifying segments and shows very significant improvement over both Cosine Similarity and Burrows' Delta. When comparing segments from the same author (Table 3.2), Jensen-Shannon Similarity with extra poetry words also does the best job of identifying segments and again shows very significant improvement over both Cosine Similarity, and Burrows' Delta. In both cases, Jensen-Shannon Similarity with extra poetry words performs the best with high significance ($p < 0.01$).

One might argue that Jensen-Shannon Similarity has an advantage because it uses plus-one smoothing and the number of non-top words, but providing the other similarity metrics with plus-one smoothing, the number of non-top words, or both does not lead to a statistically significant improvement on this evaluation for any of the similarity metrics.

It appears from these results that Jensen-Shannon Similarity is the best metric by this evaluation, so we examine its mistakes in more detail. Of the 101 segments judged to be closest to a different author, the median first occurrence of a segment by the same author was the 5th most similar segment, and the majority of

Metric	Top 145 + Top 100 in Poetry	Top 172	Top 145
Jensen-Shannon	**91.44%†‡**	91.27%†‡	90.59%†
Burrows' Delta	89.83%†	89.07%*	89.58%†
Manhattan	89.58%†	89.58%†	89.32%†
Canberra	88.05%	87.54%	87.63%
Cosine	87.37%	87.54%	87.63%

Table 3.2: How well similarity metrics based on a given set of words identify whether two segments come from the same author.
†: Results very significant ($p < 0.01$) when compared to Cosine (172).
*: Results significant ($p < 0.05$) when compared to Cosine (172).
‡: Results very significant ($p < 0.01$) when compared to Burrows' Delta (172).

the confusions are either between authors of the same time period, such as Lysias Speech 2 and Isocrates Speech 4, where both authors are Athenian Orators in the 4th Century BCE, or Theocritus' *Idylls* and Callimachus's *Hymns*, both written in the same part of the Hellenistic period. A small subset are not from similar times but have a clear reason for their similarity: for example, Arrian's *Indica* is most similar to Herodotus's *Histories* book 7, which makes sense since unlike Arrian's other work his *Indica* was modeled specifically on Herodotus — of the top 10 most similar segments, seven are from Herodotus' *Histories* and three are from Arrian's own *Anabasis*. Given that Arrian's *Indica* was *designed* to look like Herodotus more than any of his other works, it is unsurprising and perhaps even desirable that the similarity metric considers the *Indica* to look like Herodotus. Thus, even when this metric does not perfectly identify an author's identity, it does correctly identify their general signature and may incorporate some information about style beyond just author identity.

Table 3.3 shows the top 10 most similar author pairs by this metric. All of these pairs are writing at similar times and appear to have good reason to be similar.

Thus, Jensen-Shannon Similarity appears to capture that authors generally have a consistent signature and authors writing in similar genres have similar sig-

Score	Author 1	Author 2	Notes
0.9948	Demosthenes	Isaeus	"10 Attic Orators."
0.9947	Andocides	Lysias	"10 Attic Orators."
0.9944	Aelius Aristides	Dio Chrysostom	Orators, 1st/2nd century CE.
0.9940	Diodorus Siculus	Polybius	Historians, 1st/2nd century BCE.
0.9937	Euripides	Sophocles	Tragedians, 5th century BCE.
0.9936	Athenaeus	Diogenes Laertius	Writing 3̃rd century CE.
0.9933	Aeschylus	Euripides	Tragedians, 5th century BCE.
0.9932	Demosthenes	Lysias	"10 Attic Orators."
0.9931	Dionysius	Flavius Josephus	Historians from a century apart.
0.9928	Achilles Tatius	Chariton	Authors of romance novels.

Table 3.3: Top author pairs by similarity score according to Jensen-Shannon Similarity.

natures. It therefore appears to be accurately detecting actual similarity between signatures of the same author and same genre, and we can be confident that high similarities according to this metric are based on actual similiarities of the texts. We also note that Gerlach and Font-Clos independently found Jensen-Shannon Divergence to be a useful metric for comparing segments of different genres [18].

When interpreting Jensen-Shannon Similarity it is important to note that while it potentially ranges from 0 to 1, in practice the lowest similarity between two segments from the same language is 0.83, for Euclid's *Elements* Book 13 and Isocrates' Speech 21. The reason that even the lowest similarities are so high is that we always compare two texts from the same language. They are constrained by the grammar and vocabulary of their shared language and so their usage of the top words will look reasonably similar. Across our three languages (Greek, English, and Icelandic), the median Jensen-Shannon similarity for pairs of authors is roughly 0.95, so in the context of this work a 0.9 similarity score is actually very *low*.

CHAPTER 4

COMPOSITION DATE AND AUTHOR SIMILARITY

4.1 Task and Preliminary Explorations

Now that we have a similarity metric that detects similar authors, we explore whether this metric show signs of style imitation by authors: that is, how commonly Greek authors write with a signature similar to much earlier authors. To a get an initial sense of author similarity for the entire corpus, we run the analysis visible in Figure 4.1. Note that for this chapter we consider similarity based only on the top 145 words, as Icelandic, which we will consider momentarily, does not include poetic works. Each individual dot represents a pair of authors, with the number of centuries between those authors on the x-axis and the similarity of those authors on the y-axis. The centuries between two authors is an integer value, as many of these authors have only rough dates available. For ease of viewing, a small amount of random jitter is applied to the x-axis. The black line shows the average similarity for each century.

We observe a surprising degree of similarity between authors writing five to seven centuries apart, and the average similarity actually rises for these time differences when compared to authors four centuries apart. The graph appears to be bi-modal, with one cluster in the 0.95-1.00 range and one in the 0.90-0.95 range. We have also highlighted author pairs where one of the authors is Euclid or the author of the *Hymn to Dionysus*, as these two authors make up almost all of the author pairs with a similarity below 0.9. The *Hymn to Dionysus* is very short, so its normalized distribution of words is unusual, and Euclid's *Elements* is full of discussions of geometric figures that differ slightly from natural language.

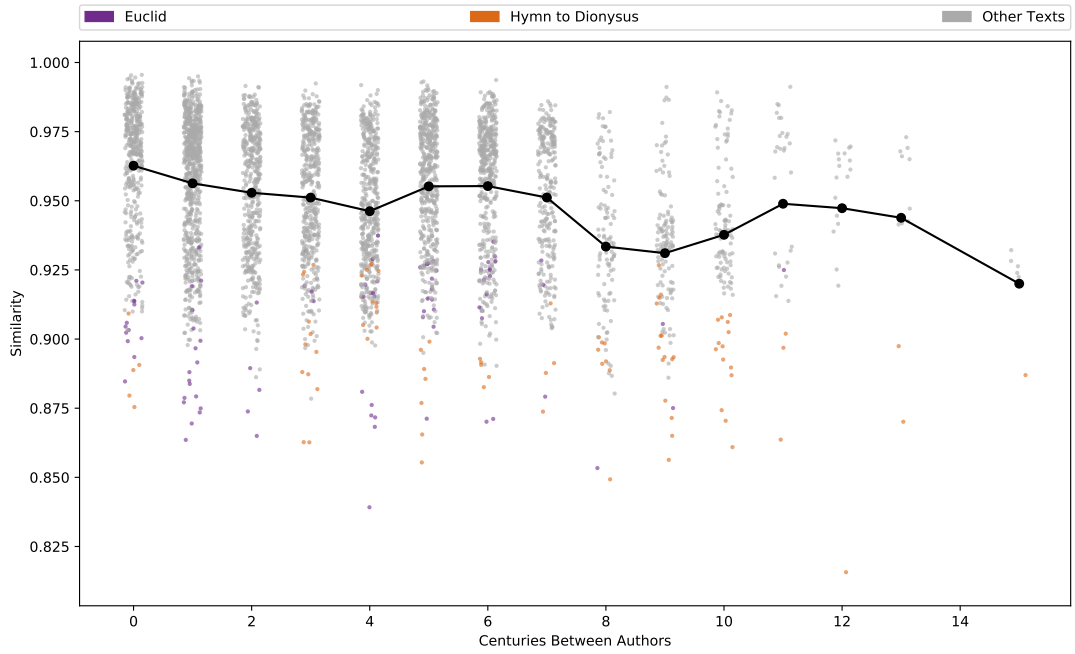

Figure 4.1: A plot of the similarity between two authors by centuries between those authors writing. The author of the *Hymn to Dionysus* (due to small sample size) and Euclid (due to unusual text structure including many geometric figures) are quite different from other authors and make up most of the lowest scores. The mean score for each century-difference is marked with the black trend line.

4.2 Ancient Greek Results

To improve on these issues, we make the following changes: we remove the *Hymn to Dionysus* and Euclid from consideration, as well as author pairs more than nine centuries apart, we color author pairs where the authors are writing in the same genre differently from those writing in a different genre, and we run a linear regression on the similarities to determine the rate of change.

Figure 4.2 makes it clear that the bi-modal behavior is explained by the genre authors are writing in: poets look more similar to poets and less similar to prose authors. There is a downward slope in similarity as authors write further apart in time, but this is mostly the result of relative frequency of prose and poetry authors over time, and only explains 3% of the variance (R2=0.03119, p=2.639e-28).

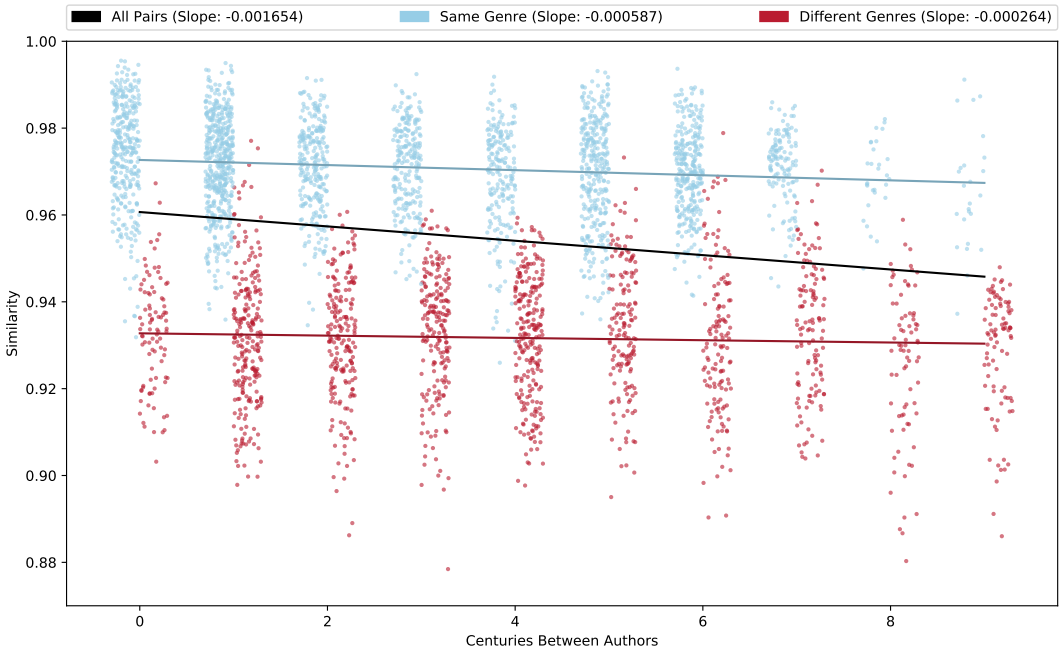

Figure 4.2: Graph of the similarity of pairs of Ancient Greek authors across different times, with authors pairs writing in the same genre marked separately from those writing in different genres.

When considering only authors writing in the same genre, or only authors writing in a different genre, the slopes are 35% and 16% as steep as the overall slope. For different genres, the century explains 0.2% of the variance and we cannot reject the hypothesis that the slope is flat (R2=2.019e-03, p=0.07592), despite the large number of points. When we consider the trend line for texts of the same genre, the slope is not zero with a p value of 3E-9, which is very small due to the number of author pairs involved. However, only 1.5% of the variance is explained by the century (R2=0.01531, p=2.985e-09). While distance in time explains very little of the variance, when the corpus as a whole is considered the matchup of genre (same genre vs different genre) accounts for almost 70% of the variation seen (R2=0.6906, p<2.0E-307).

4.3 Comparing Ancient Greek to English and Icelandic

The natural next question is whether this behavior is standard for all languages or a feature specifically of Ancient Greek. As a comparison, we include the same analysis for English authors in Figure 4.3. This shows a more intuitive distribution: authors writing further apart in time are more different, and the slope for authors of the same genre is almost 15 times steeper than for Greek: -0.0089 vs -0.0006. In addition, the century explains nearly 57% of the variance in the texts (R2=0.5688, p<2.0E-307), while, contrary to the Ancient Greek, the genre matchup explains roughly 11% (R2=0.1084, p<2.0E-307).

One might argue that English does not have the same long-term stability and standardization as Greek texts do, so we also consider Icelandic texts by prose authors from the past 900 years. All texts are standardized to Modern Icelandic

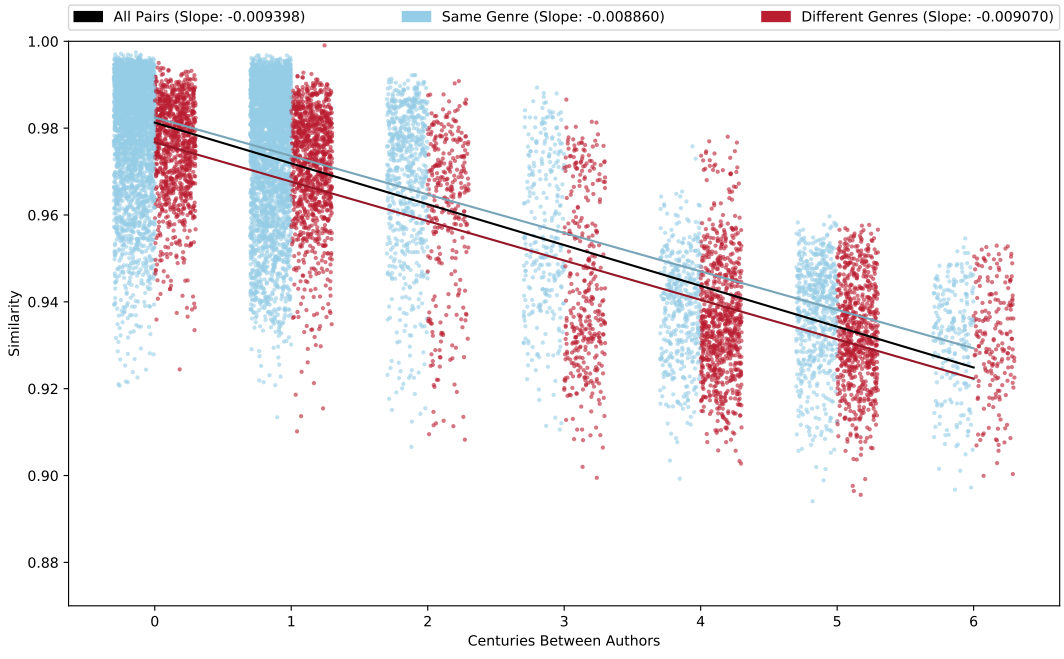

Figure 4.3: Graph of the similarity of pairs of Modern and Middle English authors across different times, with authors pairs writing in the same genre marked separately from those writing in different genres.

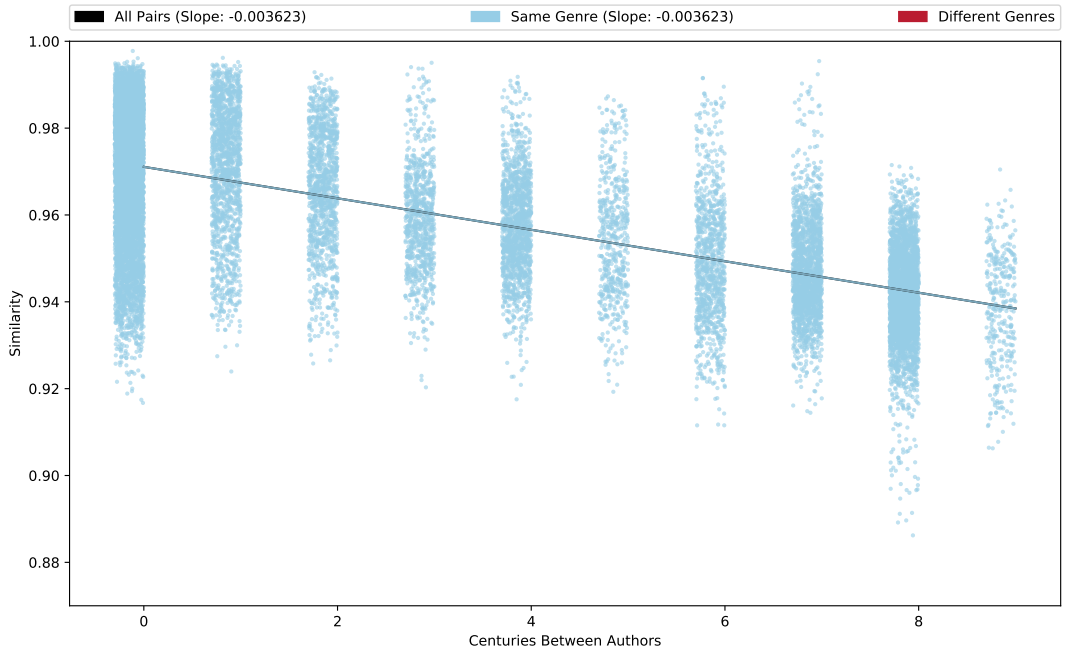

Figure 4.4: Graph of the similarity of pairs of Icelandic authors across different times. All texts are prose, so there are no instances of authors of different genres.

spelling and grammatical endings, so these texts are even more standardized than the Ancient Greek corpus. The Icelandic results are visible in Figure 4.4. We see a few author pairs four to seven centuries apart that are more similar than the bulk of authors for that time: three authors wrote sagas emulating the older sagas during the 19th and 20th centuries, so they appear rather similar to the whole range of old sagas. But these texts are a few exceptions with regard to the corpus as a whole. Like English, as the temporal distance between authors increases there is a clear decrease in similarity, though it accounts for only 44% of the variance rather than 57% ($R2=0.4360$, $p<2.0E\text{-}307$).

The English and Icelandic data does seem to support the hypothesis that English has undergone more change than Icelandic in the recent past [39]. However, the Greek texts show even *less* change than Icelandic over time, and far more instances of high similarity across a long time period. While we saw that the centuries apart explains only 3% of the variance in the Ancient Greek Corpus, it

explains over 40% for both English and Icelandic, so these two languages show clear change over time. Ancient Greek's remarkable stability when compared even to a corpus from a language that is culturally conservative and morphologically and orthographically standardized across time makes it clear that the extent of similar authors separated by many centuries is a peculiar feature of Ancient Greek.

It is also important to recognize that the Icelandic and English corpora do have some outliers. In Figure 4.4, there are a group of author pairs with a similarity above 0.97 despite writing 4 or more centuries apart. However, these pairs all include one of three later authors: Þórarinn Eldjaárn, Gísl Konráðsson, and Benedikt Gröndal, who wrote sagas in the style of much older texts. Even the English has slight outliers, including George Bernard Shaw writing plays in a (relatively) similar style to Shakespeare three centuries later. What is unusual about the Ancient Greek corpus is not necessarily that it was possible to achieve such similarity, but that there are so many very similar authors across such a long time period, with very little drop-off.

CHAPTER 5

LONG-DISTANCE SIMILARITY

5.1 Similarity of Greek Authors Four Centuries Apart

In order to get a closer look at this surprising similarity across long time periods, let us consider some of the most similar authors using our top-performing comparison metric, Jensen-Shannon Similarity with the top 145 overall words and top 100 poetry words. Of the top 100 closest author pairs, 23 pairs wrote at least four centuries apart. These pairs fall into a few clear categories:

Epic Poets: This category consists of epic poets spanning from Homer to the late Roman Empire: Apollonius similar to Homer; Oppian, Oppian of Apamea, and Tryphiodorus similar to Apollonius; and Quintus Smyrnaeus similar to both Apollonius and Homer.

Attic Style: This category consists of authors in Imperial Rome writing with signatures like those of orators and prose authors from the golden age of Athens: Aelius Aristides and Dio Chrysostom similar to Aeschines, Demosthenes, Plato, and Xenophon; Aelius similar to Andocides and Isaeus; Plutarch similar to Aeschines; and Appian and Arrian similar to Thucydides.

Christian Authors: This category consists of John of Damascus (c. 700 CE) writing with a signature similar to three prior Christian authors: Clement of Alexandria, Eusebius of Caesarea, and Basil of Caesarea.

Lastly, there is one less clear pair: Eusebius of Caesarea (c. 300 CE) and Dionysius of Halicarnassus (first century BCE). Eusebius is closer to signature of the Jewish author Flavius Josephus (the 16th most similar pair in our dataset), who wrote about a century after Dionysius with a similar signature (these two are

the 8th most similar pair in our set). So this is likely picking up Eusebius writing like Flavius Josephus, who was in turn writing in a similar manner to Dionysius of Halicarnassus. This is similar to the Attic orators above: Aelius Aristides did not attempt to imitate every one of these authors at once, but wrote in an Attic style based at times on Demosthenes, Isaeus, Plato, and Xenophon [26]. Aelius therefore appears similar to, say, Aeschines due to his similarity to Demosthenes and Isaeus, who are in turn very similar to Aeschines.

These similar pairs do not just reflect the fact that some authors writing four centuries apart are more similar than others. These pairs are in the top 100 (2.4%) most similar pairs out of 4,186 total, and of the top 2.4% of pairs of English authors not a single pair is writing more than one century apart. These authors are more similar than authors writing at the same time in similar genres, such as Isocrates and Lysias (members of the "Ten Attic Orators"), Plato and Xenophon, or Aratus and Callimachus, and they are far more similar than the historians Herodotus and Thucydides (674th most similar). So the high similarity between these author pairs is clearly signalling similar styles rather than random chance.

These similar authors are also reasonably consistent across the other metrics considered, so when compared using other metrics in Table 5.1 the pairs usually appear in the top 100 pairs and in all but three cases appear in the top 10% (418) of pairs. The three pairs not in the top 10% according to Canberra distance are very similar by Manhattan distance, which shows the limitation of Canberra distance: if one author uses a word six times out of two hundred thousand words and the other uses it once out of a hundred thousand, Canberra distance considers those authors to be very different. As discussed above, Jensen-Shannon Similarity provides a compromise between the the absolute and relative distances captured by Manhattan and Canberra similarity.

	Rank according to				
	Jensen-Shannon	Burrows' Delta	Manhattan	Canberra	Cosine
Aelius Aristides, Demosthenes	11	20	6	14	9
Apollonius, Quintus Smyrnaeus	20	69	86	84	29
Apollonius, Homer	23	56	45	253	6
Clement, John of Damascus	34	32	66	102	39
Dionysius, Eusebius	35	38	41	57	81
Apollonius, Oppian	37	255	222	447	56
Aeschines, Plutarch	40	33	27	44	41
Aelius Aristides, Plato	41	46	29	10	48
Demosthenes, Dio Chrysostom	44	117	35	24	78
Apollonius, Oppian of Apamea	46	417	340	1027	169
Aeschines, Aelius Aristides	47	86	50	29	67
Aelius Aristides, Isaeus	53	124	38	62	97
Appian, Thucydides	66	26	47	91	31
Eusebius, John of Damascus	69	105	100	138	51
Andocides, Aelius Aristides	70	125	83	66	151
Aeschines, Dio Chrysostom	71	148	82	55	215
Apollonius, Tryphiodorus	72	516	636	1462	37
Arrian, Thucydides	75	53	18	213	12
Basil, John of Damascus	79	83	89	32	401
Aelius Aristides, Xenophon	84	74	49	34	167
Homer, Quintus Smyrnaeus	91	222	294	344	63
Dio Chrysostom, Xenophon	96	153	124	50	264
Dio Chrysostom, Plato	98	150	102	26	113

Table 5.1: Rank of highly similar authors writing at least four centuries apart by different metrics.

Such close similarities across such a long time period does not necessarily imply active imitation, but in combination with other information it does provide evidence of successful imitation. For example, as discussed above, Arrian's *Indica* is allegedly modeled after the work of Herodotus, and in fact for all nine books of Herodotus' *Histories*, the *Indica* is the most similar segment after the other books of the Histories. Similarly, epic poets after Homer were all attempting to write

like him, so the close similarity between Apollonius and Homer implies successful imitation on the part of Apollonius.

However, impressive similarity between authors brings us to our second question: how exactly are authors achieving this similarity? The similarity score alone does not readily provide this information, so we perform further analyses.

5.2 Analyzing How Temporally Distant Authors Achieve Similarity

It is now worth examining what is happening in top pairs from some of these categories. Our first key question is whether this similarity is the result of strategic use of a few words, or a wider imitation across words. There are two broad hypotheses for how authors have writing signatures similar to models from many centuries earlier:

A. These signatures are similar due to very similar usage of a small number of words.

B. These signatures are similar due to using many or most words in very similar fashion.

To get an intuition for which of these hypotheses is supported by the data, we consider how similar authors appear when we ignore their most similar words (that is, the words these two authors use with the most similar frequency). If hypothesis A is correct and the similarity comes from a few similar words, ignoring their most similar words should make the most similar pairs look more like the average pair. Figure 5.1 shows the similarity between pairs of authors writing four

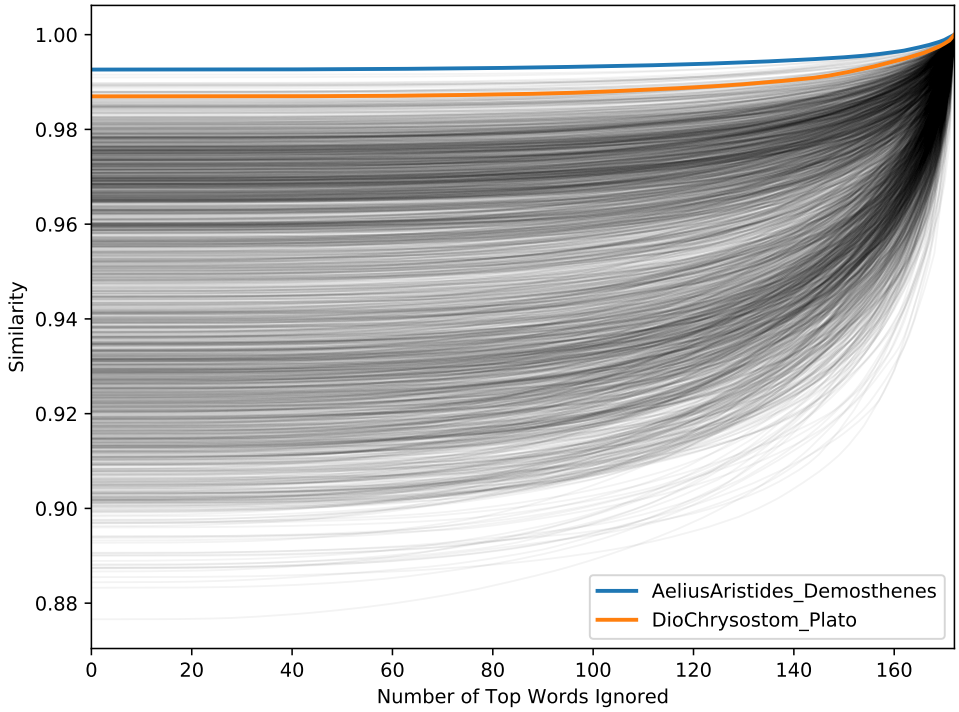

Figure 5.1: Visualization of author similarity when ignoring the `x` words with the closest frequencies (with `x` on the x-axis). Only author pairs writing at least four centuries apart are shown. The best and worst pair from Table 5.1 are highlighted to show the range of these top pairs.

centuries apart when we ignore their `x` most similar words. When `x = 0`, we see the Jensen-Shannon Similarity of each author; when `x = 50`, we see the Jensen-Shannon Similarity of each author ignoring the 50 words they use with most similar frequency, and so on.

If hypothesis A were correct, and the high similarity was the result of using, say, 20 words very similarly, we would expect the lines corresponding to very similar pairs to be indistinguishable from less similar pairs once these top 20 words were ignored. Instead, we see that no matter how many words are ignored, similar author pairs still appear very similar compared to other pairs, which supports the idea that the similarity in author signatures is the result of using most words in a

very similar fashion: hypothesis B above.

A closer view of the word usage backs up this finding. Considering the most similarly used word, τῶν, of our most similar long-term pair, Aelius Aristides and Demosthenes, the two authors use the word with almost identical frequency. However, the genitive plural nouns paired with τῶν are not used with identical frequencies, so similar τῶν frequency is not capturing identical usage of entire phrases. It is rather that Aelius is using the article with genitive plurals in a similar manner to Demosthenes, and because Aelius is using many words reasonably similarly, in some cases the frequencies happen to match nearly exactly.

5.3 Word Usage by Temporally Distant Authors

Instead of looking at over a hundred words individually, one way to get an intuitive sense for how these authors are writing is to group words into a small number of categories. The first set of categories we examine is dividing words by part of speech. We show the resulting author signatures in Figure 5.2 for four pairs of authors that show similarity across many centuries:

- Speeches: Aelius Aristides, writing five centuries after Demosthenes.
- Epic Poetry: Apollonius Rhodius, writing roughly five centuries after Homer.
- Christian Prose: John of Damascus, writing five centuries after Clement of Alexandria.
- History: Appian, writing six centuries after Thucydides.

These charts help give a sense of what common words the authors are using. Apollonius's similarity to Homer is driven by far more frequent use of particles and far

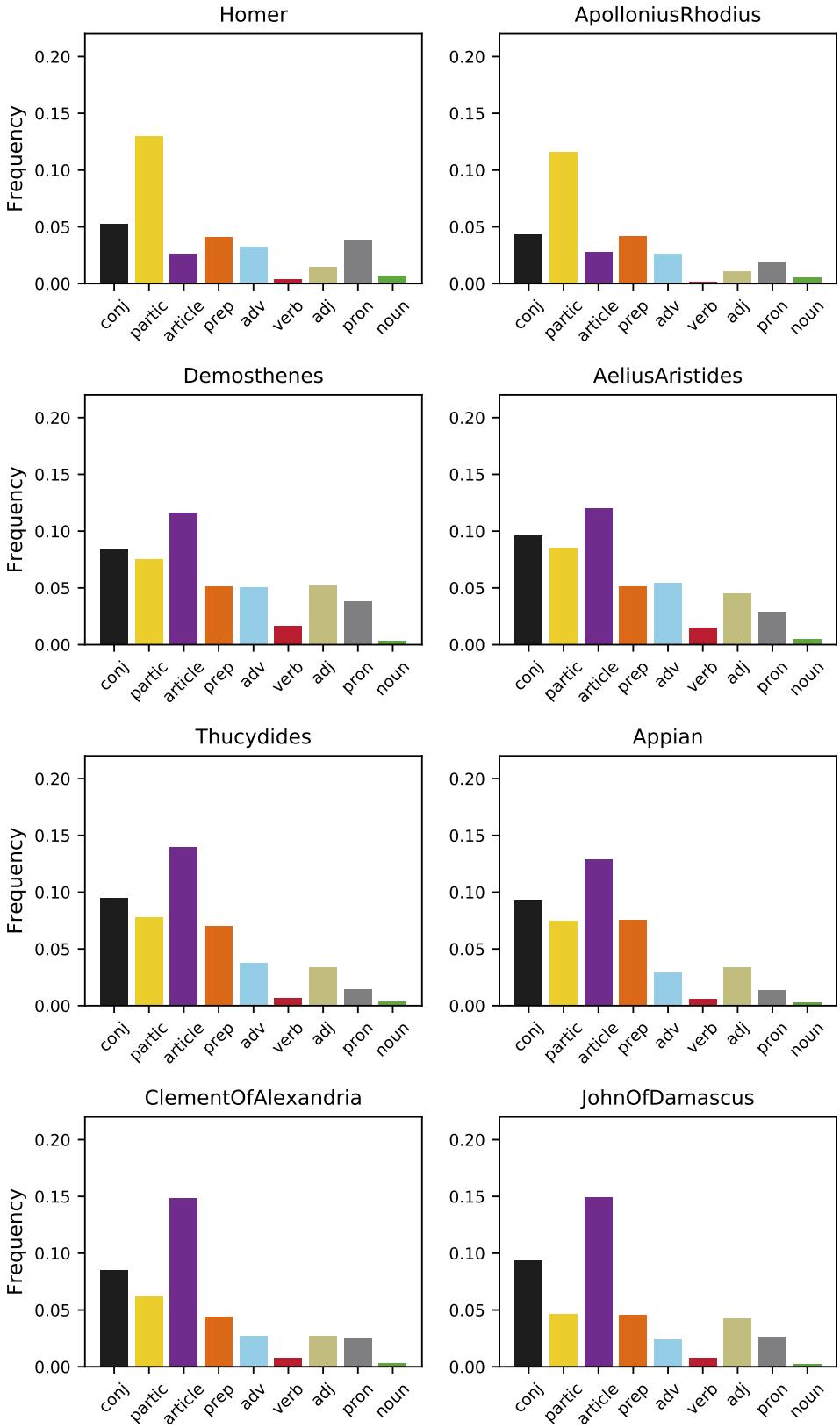

Figure 5.2: Relative frequencies of each part of speech in 8 authors, chosen to illustrate pairs of authors with similar signatures over at least four centuries. The later author is on the right.

less use of conjunctions and the article when compared to the prose authors. When compared to the orators, Appian's more frequent use of prepositions and less frequent use of pronouns gives him a "Thucydidean" signature; this also makes sense as one would expect pronouns to appear more frequently in speeches than history. These charts also make clear that the similarity of these authors occurs across all categories seen, rather than showing similarity in just one or two categories.

In addition to grouping words by their part of speech, we also consider grouping words by the authors who use them, which allows us to get categories with more distinction between authors. We place all words in a high-dimensional space defined by the frequency with which each author uses that word, then run a K-Means algorithm to group similar words into nine categories.[1] A two-dimensional projection of this high dimensional space is visible in Figure 5.3. Due to the sharp distinction between poetry and prose in these texts, most word groups mainly distinguish between those two categories. καί and δέ are the two most frequent words and thus get their own categories, the first used more in prose and the second more in poetry. The τῶν and τό groups contain forms of the relative pronoun which are relatively infrequent in poetry and more frequent in prose, but the τῶν words show a sharp distinction between prose and poetry while the τό group words are very infrequent in epic poetry, slightly more frequent in other poetry, and much more frequent in prose. On the whole poetic authors use words in the μέν group less frequently, but there is less of a clear division as Pindar is 24th out of 92 and Aratus and Oppian are roughly in the middle. The τε group contains this single word, which seems to be slightly more common in older texts and poetry. The τούς group has function words more common in prose, while the ἐς group has function words and seems to be capturing texts that are in the Ionic dialect (note the more

[1] We found nine to be the largest number of categories that were reasonably stable across different runs.

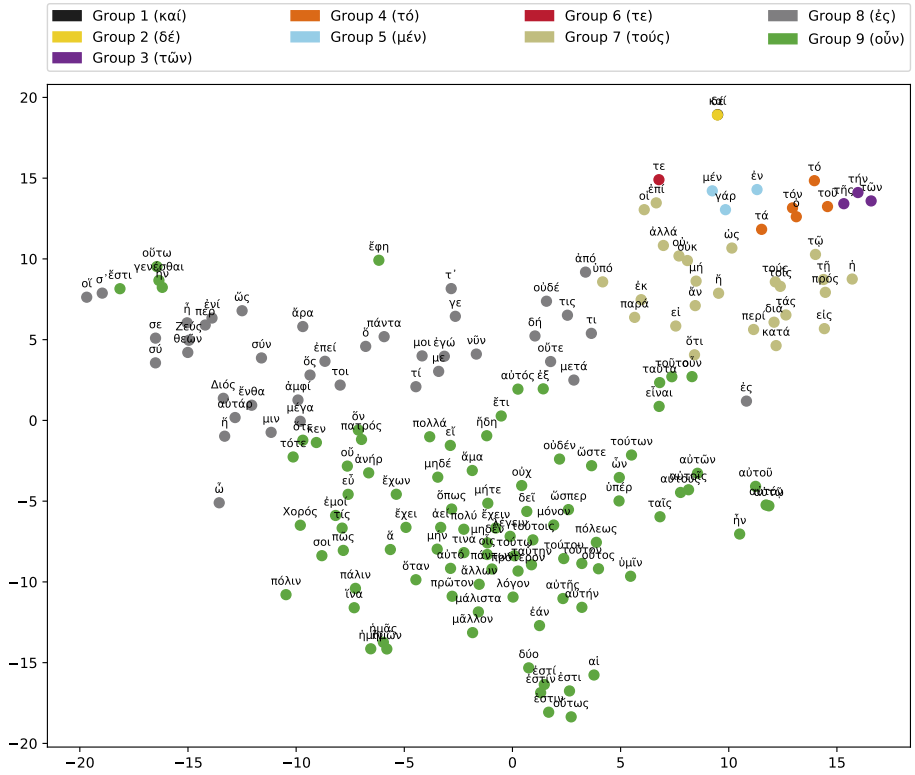

Figure 5.3: tSNE visualization of words grouped by which segments they occur in. Words that are closer together are used with similar frequency by different authors.

common form of ἐς, this category's most frequent word, is εἰς, which appears in Group 7 - τούς). The οὖν group is miscellaneous assortment of the remaining words, but most of them are function words that appear less often in poetry and more often in Attic rhetoric.

Once again, we see in Figure 5.4 that texts by similar authors have similar signatures. This word grouping also shows slightly more differentiation in author signature between different authors: for example, in Figure 5.2 Thucydides and Clement show slight differences across many categories, while in Figure 5.4 they show clear difference in the τό and τε group. This makes sense as these words groups were chosen to differentiate authors.

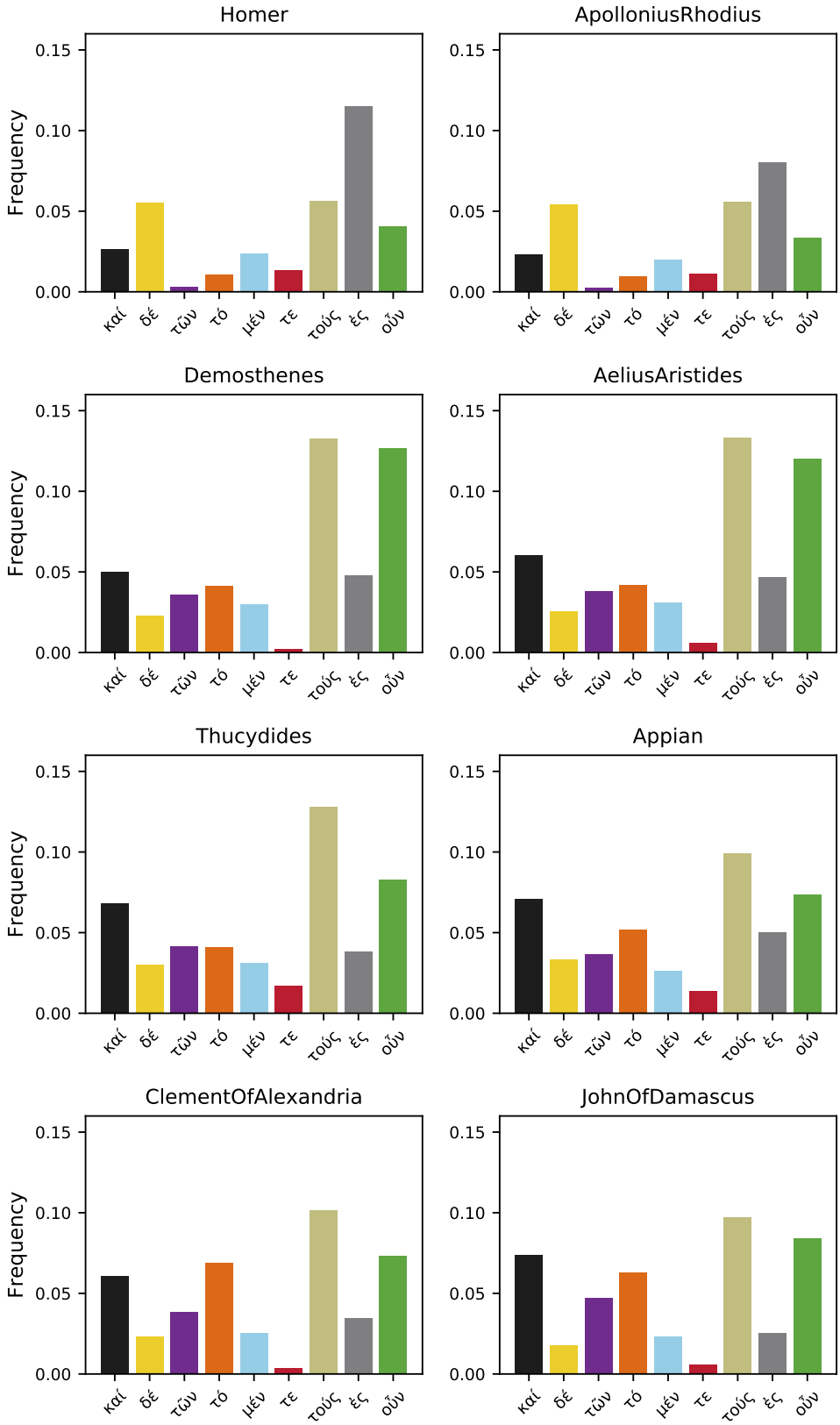

Figure 5.4: Relative frequencies of chosen word groups in 8 authors, chosen to illustrate pairs of authors with similar signatures over at least four centuries. The later author is on the right.

CHAPTER 6

CONCLUSION

Having examined Ancient Greek author similarity in terms of what is happening and how it is happening, we now turn to reasons *why* this similarity may be present. The comparison to English and Icelandic makes it clear that there is something unusual going on with this similarity, beyond what one might expect for even a conservative language like Icelandic. In fact, many of the later authors would have spoken the Koine (common) dialect of Greek rather than the Attic, Ionic, or Homeric dialects of the texts they imitated. When Callimachus wrote in the Homeric dialect, which combines features of Ionic and Aeolic, both dialects had been mostly replaced by the Attic-based Koine, and by the time of Aelius Aristides and his contemporaries "those who wanted to write the best Attic... clearly needed help... no one had spoken the prescribed model Attic for centuries" [22]. The similarity between authors across long time periods was *not* due to a similarity in the spoken language across those time periods, but rather to the authors' usage of older forms of Greek no longer in common use.

There are three major factors that might explain the similarity or difference between texts: their genre, dialect, and time period. In both the English and Icelandic cases distance in time correlates with difference between texts, but this is not the case with Ancient Greek to a surprising degree. Instead, in the Ancient Greek case, genre has a much more significant influence on the similarity of texts: prose works from six centuries apart are usually more similar than a prose-poetry pair written in the same century. Part of this explanation is that works of poetry had additional constraints that led to different word usage from prose: for example, the constraints of meter made it more difficult to use the article with every noun. Another likely cause of this is the connection between genre and dialect. In

the classical period different genres were generally associated with specific dialects, so authors wrote lyric poetry in the associated dialect, literary Doric, regardless of their own native dialect [51]. Later authors, especially orators and epic poets, seem to have followed this model and intentionally written in older dialects. It is even possible that this remarkable similarity is *caused* by the fact that the later authors were writing in a non-native dialect of the language. If Aelius Aristides spoke Koine Greek but wanted his speeches to sound like Demosthenes, he had to learn Attic Greek *from* earlier authors like Demosthenes; there were no native Attic speakers to teach the language. Since he would therefore be learning Demosthenian specifically rather than Attic as a whole, it makes sense that his speeches may look more similar to Demosthenes than a generic Attic text. There is in fact evidence from ancient authors that one of the suggested ways for developing and improving one's own style was to memorize entire texts by heart, which would make the similarity in style more understandable [35, 11].

Cultural considerations are also important to understanding how this happened: similarity to older models was the culturally correct way to write epic poems, speeches, high-register histories, and so on, so authors would have had a cultural incentive to do a good job at this, since a work that looked like the older models would be more well-regarded. There may also be survivor bias in the corpus: the works that survive to today are, for the most part, the best works by the best authors, according to cultural standards that viewed writing in certain styles as "the best." If the only Greek speeches from the Roman Empire preserved to the modern day are the ones that are "the best," where the best is defined as the most like speeches from Classical Athens, the strength of this effect across the corpus makes slightly more sense.

While we see these similarities to an unusual degree in the Ancient Greek corpus, this is likely not a capability that only Ancient Greek humans had. Modern research has shown that author signatures are not immutable: there are examples of authors varying their own signature in different works or even within a single book [15, 41, 9]. However, there is less modern work on how an author might adjust their signature to be more like a specific model: the ancient sources suggest copying and memorization would help, but this hypothesis has not been proven [11, pages 12-13]. One future direction of exploration is examining in more detail the mechanisms for achieving this similarity. As we discussed above, there was not an exact correspondence between usage of individual words, but whether the broadly similar word usage come from sentence and clause construction would be an interesting avenue of further research, both from a computational and traditional classics perspective. Another interesting question would be whether modern individuals could achieve this sort of similarity as well. Asking a group of actors who have memorized plays of Shakespeare and a control group of non-actors to each write a text sample in the style of Shakespeare would be an interesting experiment in whether memorization leads to better imitation.

One might also try to examine or isolate the impact of scribal copying by analyzing and comparing different surviving manuscripts to determine their impact on these types of analyses. If scribes are able to add their own signature to a text, as work on Dutch manuscripts suggest [29, 55], how does that impact our results and how might the effect be handled? In addition, comparison to further languages may yield additional interesting results. There are many more languages with multi-century literary traditions and the potential for imitation that could be compared and contrasted to the Ancient Greek system and help clarify how much of an outlier this tradition is.

And while this work suggests a variety of further areas of exploration, it conclusively shows that Ancient Greek authors of the Hellenistic and Roman periods wrote in a remarkably similar fashion to the predecessors in the classical period, at least based on their usage of common function words. Comparisons to English and Icelandic show that this is not a natural feature of every language, and analysis of individual word usage shows this similarity happens across a broad spectrum of words rather than a select few. While we approached this problem from a different direction than the usual Classics approach, we hope this too is instructive. As is increasingly being recognized, computational techniques are able to supplement and work alongside more traditional methods in Classics scholarship, providing useful context, answering questions in different ways, and opening new doors for further study of classical texts.

BIBLIOGRAPHY

[1] Włodzimierz Appel. Die homerischen hapax legomena bei Quintus Smyrnaeus: Adverbien. *Glotta*, 71(3./4. H):178–188, 1993.

[2] Kristján Árnason. *The phonology of Icelandic and Faroese*. Oxford University Press, 2011.

[3] Simon Barnes-Sadler. A digital humanities approach to inter-Korean linguistic divergence: Stylometric analysis of ROK and DPRK journalistic texts. *S/N Korean Humanities*, 4(1):127–153, 2018.

[4] Francesca Benatti and Justin Tonra. English bards and unknown reviewers: a stylometric analysis of Thomas Moore and the Christabel Review. *Breac: A Digital Journal of Irish Studies*, 2015.

[5] José Nilo G Binongo. Who wrote the 15th book of Oz? an application of multivariate analysis to authorship attribution. *Chance*, 16(2):9–17, 2003.

[6] Eleni Bozia. Measuring tradition, imitation, and simplicity: The case of Attic oratory. *Corpus-Based Research in the Humanities (CRH)*, page 23, 2015.

[7] Eleni Bozia. Atticism: The language of 5th-century oratory or a quantifiable stylistic phenomenon? *Open Linguistics*, 2(1), 2016.

[8] John Burrows. 'Delta': a measure of stylistic difference and a guide to likely authorship. *Literary and linguistic computing*, 17(3):267–287, 2002.

[9] John Burrows. Who wrote Shamela? verifying the authorship of a parodic text. *Literary and Linguistic Computing*, 20(4):437–450, 2005.

[10] Geoffrey Chaucer. *Chaucer's Works, Volume 4 (of 7) — The Canterbury Tales*. University of Oxford, 2007.

[11] Donald Lemen Clark. Imitation: Theory and practice in Roman rhetoric. *Quarterly Journal of Speech*, 37(1):11–22, 1951.

[12] Neil Coffee, Jean-Pierre Koenig, Shakthi Poornima, Christopher W Forstall, Roelant Ossewaarde, and Sarah L Jacobson. The Tesserae Project: intertextual analysis of Latin poetry. *Literary and linguistic computing*, 28(2):221–228, 2012.

[13] Dionysius. *The Critical Essays.* Harvard University Press., 2015.

[14] Maciej Eder et al. Style-markers in authorship attribution a cross-language study of the authorial fingerprint. *Studies in Polish Linguistics*, 6(1):99–114, 2011.

[15] Mel A Evans. Tudor women writing: Multimodal style and identity in the english letters and prose of Queen Katherine Parr and Princess Elizabeth. *Studies in Variation, Contacts and Change in English*, 17, 2016.

[16] Christopher Forstall, Neil Coffee, Thomas Buck, Katherine Roache, and Sarah Jacobson. Modeling the scholars: Detecting intertextuality through enhanced word-level n-gram matching. *Digital Scholarship in the Humanities*, 30(4):503–515, 2014.

[17] Finnur Friðriksson. *Language change vs. stability in conservative language communities. A case study of Icelandic.* Institutionen för lingvistik, 2008.

[18] Martin Gerlach and Francesc Font-Clos. A standardized Project Gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *arXiv preprint arXiv:1812.08092*, 2018.

[19] Vivienne J Gray. Xenophon's Cynegeticus. *Hermes*, 113(H. 2):156–172, 1985.

[20] NGL Hammond. The speeches in Arrian's Indica and Anabasis. *The Classical Quarterly*, 49(1):238–253, 1999.

[21] Sigrún Helgadóttir, Ásta Svavarsdóttir, Eiríkur Rögnvaldsson, Kristín Bjarnadóttir, and Hrafn Loftsson. The tagged Icelandic corpus (MÍM). *Language Technology for Normalisation of Less-Resourced Languages*, page 67, 2012.

[22] Geoffrey Horrocks. *Greek: A History of the Language and its Speakers.* John Wiley & Sons, second edition, 2010.

[23] John D Hunter. Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(3):90, 2007.

[24] George Norlin Isocrates and Larue Van Hook. *Isocrates.* Harvard University Press., 2015.

[25] Einar Freyr Sigurðsson Joel C. Wallenberg, Anton Karl Ingason and Eiríkur Rögnvaldsson. Icelandic parsed historical corpus version 0.9, 2011.

[26] Christopher P Jones. Aelius Aristides, ΕΙΣ ΒΑΣΙΛΕΑ. *The Journal of Roman Studies*, 62:134–152, 1972.

[27] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001. [Online].

[28] Konstantinos A Kapparis. *Apollodoros "Against Neaira"[D 59]*, volume 53. Walter de Gruyter, 2012.

[29] Mike Kestemont, Van Dalen-Oskam, et al. Predicting the past: memory-based copyist and author discrimination in medieval epics. In *Proceedings of the twenty-first Benelux conference on artificial intelligence (BNAIC 2009)*, volume 21, pages 121–128, 2009.

[30] Mike Kestemont, Justin Stover, Moshe Koppel, Folgert Karsdorp, and Walter Daelemans. Authenticating the writings of Julius Caesar. *Expert Systems with Applications*, 63:86–96, 2016.

[31] Thomas Koentges. Computational Analysis of the Corpus Platonicum. Technical report, Center for Hellenic Studies, Harvard University, 04 2018.

[32] Michael Martin Kumpf. *The Homeric hapax legomena and their literary use by later authors, especially Euripides and Apollonius Rhodius*. PhD thesis, The Ohio State University, 1975.

[33] Dominique Labbé. Experiments on authorship attribution by intertextual distance in English. *Journal of Quantitative Linguistics*, 14(1):33–80, 2007.

[34] Shibamouli Lahiri. Complexity of Word Collocation Networks: A Preliminary Structural Analysis. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 96–105, Gothenburg, Sweden, April 2014. Association for Computational Linguistics.

[35] Lucian. *How to Write History*. Harvard University Press, 1959.

[36] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[37] Thomas Malory. *Le Morte Darthur*. University of Michigan Humanities Text Initiative, 1997. Provided by the University of Michigan's Corpus of Middle English Prose and Verse.

[38] Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze. Introduction to information retrieval. *Natural Language Engineering*, 16(1):100–103, 2010.

[39] James Milroy and Lesley Milroy. Linguistic change, social network and speaker innovation. *Journal of linguistics*, 21(2):339–384, 1985.

[40] Haukur Þorgeirsson. How similar are Heimskringla and Egils saga? an application of Burrows' delta to Icelandic texts. *European Journal of Scandinavian Studies*, 48(1):1–18, 2018.

[41] Lisa Pearl, Kristine Lu, and Anousheh Haghighi. The character in the letter: Epistolary attribution in Samuel Richardson's Clarissa. *Digital Scholarship in the Humanities*, 32(2):355–376, 2016.

[42] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[43] Antonios Rengakos. Apollonius Rhodius as a Homeric scholar. In *Greek Literature in the Hellenistic Period*, pages 241–264. Routledge, 2018.

[44] Eiríkur Rögnvaldsson and Sigrún Helgadóttir. Morphosyntactic tagging of Old Icelandic texts and its use in studying syntactic variation and change. In *Language Technology for Cultural Heritage*, pages 63–76. Springer, 2011.

[45] Wilhelm Schmid. *Der Atticismus in seinen Hauptvertretern von Dionysius von Halikarnass bis auf den zweiten Philostratus*, volume 4. W. Kohlhammer, 1896.

[46] Mike Scott. Shakespeare corpus. Accessed August 14th, 2018.

[47] Skipper Seabold and Josef Perktold. Statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

[48] David A Smith, Jeffrey A Rydberg-Cox, and Gregory R Crane. The Perseus Project: A digital library for the humanities. *Literary and Linguistic Computing*, 15(1):15–25, 2000.

[49] Justin Stover and Mike Kestemont. Reassessing the Apuleian corpus: a com-

putational approach to authenticity. *The Classical Quarterly*, 66(2):645–672, 2016.

[50] TEAMS Middle English text series. Accessed August 16th, 2018.

[51] Olga Tribulato. Literary dialects. *A Companion to the Ancient Greek Language*, page 388, 2010.

[52] Stavros Tsitsiridis. *Platons Menexenos: einleitung, text und kommentar*, volume 107. Walter de Gruyter, 2011.

[53] Karina van Dalen-Oskam. The secret life of scribes. exploring fifteen manuscripts of Jacob van Maerlant's Scolastica (1271). *Literary and linguistic computing*, 27(4):355–372, 2012.

[54] Karina van Dalen-Oskam. Epistolary voices. the case of Elisabeth Wolff and Agatha Deken. *Literary and Linguistic Computing*, 29(3):443–451, 2014.

[55] Karina van Dalen-Oskam and Joris Van Zundert. Delta for middle Dutch—author and copyist distinction in Walewein. *Literary and Linguistic Computing*, 22(3):345–362, 2007.

[56] Stefan Van Der Walt, S Chris Colbert, and Gael Varoquaux. The numpy array: a structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2):22, 2011.

[57] Ying Zhao, Justin Zobel, and Phil Vines. Using relative entropy for authorship attribution. In *Asia Information Retrieval Symposium*, pages 92–105. Springer, 2006.

APPENDIX A

BRIEF EXPLANATION OF MACHINE LEARNING ALGORITHMS IN SECTION 2.3

In Section 2.3.2, we use three machine learning algorithms which may be unfamiliar to some audiences. These classifiers are trained on some portion of data, called the *training set*, and then predict the labels of other data points, called the *test set*. We briefly describe each of the algorithms used here.

Majority Class Classifier

This classifier always chooses the class that was most common in the training set. So if the training set had 30 poetry segments and 33 prose segments, the majority class is prose and this classifier would label all texts in the test set as prose. This algorithm is not particularly good, but it provides a useful baseline for how a very simple algorithm would perform.

K Nearest Neighbors

For each test point, this algorithm find the k closest points in the training set and picks the most common label from those points, where k is an integer. The intuition is that a point will have a similar label to the points close to it, where "close" is defined as the distance between the points in the high dimensional space defined by the feature vectors.

Multinomial Naive Bayes

The Multinomial Naive Bayes classifier predicts what it thinks is the most likely (i.e. highest probability) category c for a test point t. It determines the probability that t has category c based on the probability of seeing category c (more common categories are more likely) and the probability that an author or segment in category c would produce the words seen in t. It calculates these probabilities using the training set. For a more complete description, see [38, Section 13.2].

APPENDIX B

CODE

The code for this work, along with instructions for acquiring the corpora used, is available to view or download at `https://github.com/storey/masters_thesis`.

This work uses the packages `scipy` [27], `numpy` [56], `scikitlearn` [42], and `statsmodels` [47] for data processing and analysis and `matplotlib` [23] for charts.

APPENDIX C

TOP WORDS

See Table C.1.

Token	A	P	Token	A	P	Token	A	P	Token	A	P
καί	1	2	δή	44	31	τότε	87	79	ὅταν	130	
δέ	2	1	τοῦτο	45	83	ἐστι	88		τινα	131	
τῶν	3	39	ἀπό	46	44	πόλιν	89		πολύ	132	
τήν	4	40	οὐδέ	47	15	ἡμῖν	90		ἡμᾶς	133	
τό	5	17	μετά	48	58	οὐχ	91		γενέσθαι	134	
μέν	6	5	αὐτόν	49		τ᾽	92	8	ἡμῶν	135	
τοῦ	7	54	τι	50	47	μόνον	93		αὐτῆς	136	
τῆς	8	72	ἦν	51	90	πολλά	94	97	ἀεί	137	
τόν	9	11	γε	52	12	πρῶτον	95		ἦν	138	
ἐν	10	7	αὐτῷ	53		δεῖ	96		λόγον	139	
γάρ	11	3	αὐτοῦ	54		τούτου	97		πόλεως	140	
τε	12	4	τις	55	34	ἐστίν	98		λέγειν	141	
ὁ	13	23	αὐτῶν	56		ἅμα	99		τούτῳ	142	
τά	14	37	οὔτε	57	69	μηδέ	100		ταύτην	143	
τούς	15	53	ὦ	58	18	αὐτήν	101		ὅν	144	
τοῖς	16	71	ταῖς	59		μήτε	102		μηδέν	145	
πρός	17	38	τούτων	60		ἄ	103		ὡς		25
ἐπί	18	9	αὐτοῖς	61		ἵνα	104		χορός		26
τῷ	19	62	τί	62	30	με	105	22	τοι		50
οἱ	20	16	ἔτι	63	56	τούτοις	106		μιν		51
ὡς	21	13	νῦν	64	43	σύ	107	36	σε		52
ἀλλά	22	6	οὐδέν	65		οἵ	108	55	ἤ		57
ἤ	23	28	ἐξ	66	75	ἄλλων	109		ἀμφί		63
κατά	24	35	ὧν	67		ὅς	110	48	αὐτάρ		64
εἰς	25	41	ὥστε	68		πάντων	111		διός		65
μή	26	14	αὐτός	69	77	μήν	112		τίς		66
ἄν	27	24	ὅ	70	46	πάλιν	113		σ᾽		68
περί	28	60	ἐγώ	71	29	ἐστί	114		ἐνί		70
οὐ	29	10	μοι	72	27	σοι	115	84	σύν		73
τῇ	30		ὥσπερ	73		ἔστι	116		περ		76
ἡ	31	67	πάντα	74	59	οὗτος	117		μέγα		78
τάς	32		αὐτούς	75		ἔχει	118		θεῶν		80
διά	33	61	ἐστιν	76		ὑμῖν	119		ὅτε		81
οὐκ	34	21	οὕτως	77		ἐάν	120		ἐμοί		82
ἐκ	35	33	ἄρα	78	20	εἴ	121	88	οὗ		85
ὅτι	36	86	μᾶλλον	79		πῶς	122	100	ἔνθα		89
ἐς	37	19	ὑπέρ	80		δύο	123		ἦ		91
ὑπό	38	42	αἰ	81		τοῦτον	124		ἔχων		92
οὖν	39	87	ἔφη	82		οἷς	125		εὖ		93
εἶναι	40		ἤδη	83	94	ἔχειν	126		ἀνήρ		95
εἰ	41	32	ἐπεί	84	45	ὅπως	127		κεν		96
παρά	42	49	οὕτω	85		πρότερον	128		ζεύς		98
ταῦτα	43	74	μάλιστα	86		αὐτό	129		πατρός		99

Table C.1: List of tokens used and their rank in the top 145 tokens found in all texts (**A**) and rank in the top 100 tokens found in poetry texts (**P**).